Likelihood ratio methods for forensic comparison of evaporated gasoline residues

P. Vergeer, A. Bolck, L.J.C. Peschier, C.E.H. Berger, J.N. Hendrikse

Netherlands Forensic Institute, P.O. Box 24044,
2490 AA The Hague, The Netherlands

In the investigation of arson, evidence connecting a suspect to the fire scene may be obtained by comparing the composition of ignitable liquid residues found at the crime scene to ignitable liquids found in possession of the suspect. Interpreting the result of such a comparison is hampered by processes at the crime scene that result in evaporation, matrix interference, and microbial degradation of the ignitable liquid.

Most commonly, gasoline is used as a fire accelerant in arson. In the current scientific literature on gasoline comparison, classification studies are reported for unevaporated and evaporated gasoline residues. In these studies the goal is to discriminate between samples of several sources of gasoline, based on a chemical analysis. While in classification studies the focus is on discrimination of gasolines, for forensic purposes a likelihood ratio approach is more relevant.

In this work, a first step is made towards the ultimate goal of obtaining numerical values for the strength of evidence for the inference of identity of source in gasoline comparisons. Three likelihood ratio methods are presented for the comparison of evaporated gasoline residues (up to 75% weight loss under laboratory conditions). Two methods based on distance functions and one multivariate method were developed. The performance of the three methods is characterized by rates of misleading evidence, an analysis of the calibration and an information theoretical analysis.

The three methods show strong improvement of discrimination as compared with a completely uninformative method. The two distance functions perform better than the multivariate method, in terms of discrimination and rates of misleading evidence.

Keywords: Chromatography, Evidence evaluation, Distances functions, Gasoline comparison, Multivariate distributions

1. Introduction

At the Netherlands Forensic Institute, procedures for forensic gasoline (residue) comparison have developed from concluding in terms of the probability of same source or different source to concluding in terms of a likelihood ratio (LR) as a measure for the strength of evidence. A LR is assigned based on the comparison of a chromatographic analysis of a sample containing gasoline traces found at a crime scene and a sample containing gasoline traces found in connection to a suspect.

A likelihood ratio is defined as the ratio of the probability of the evidence given each of two competing hypotheses (for references in the forensic literature, see [1-7]). For example, it is reported that the observed (dis)similarities in the chromatograms are much more probable when the gasolines share a common source than when they are from different sources (the sources will be defined later). In forensic science, as a matter of convention, the prosecution hypothesis (here: same source) features in the numerator of the LR, while the defense hypothesis (here: different source) features in the denominator of the LR.

The present study is part of a research program to develop computer-based methods for forensic gasoline comparison resulting in a numerical LR. Computer-based methods may assist expert judgment by providing LRs that are transparent and have a clear empirical foundation in the training databases used. In this program a number of steps will need to be taken before the methods are suitable for application in forensic casework. In the current paper, evaporated gasoline residues are compared with other evaporated gasoline residues and unevaporated gasolines. Taking into account evaporation is a first step towards application in casework. In future steps the influence of the matrix and microbial degradation of gasoline residues found at crime scenes [8] will need to be taken into account.

1.1 The likelihood ratio framework for evidence evaluation

Presenting the strength of the evidence as a likelihood ratio is in concordance with the role of an expert witness in court, and leaves room for the other actors (judge, jury and other witnesses) to make their contributions. The general framework for application of a LR revolves around Bayes' theorem. When applied in a forensic setting, it dictates a connection between the odds of the two competing hypotheses (the prosecution and the defense hypothesis) prior and posterior to the taking into account of new evidence, and the strength of that new evidence as given by a LR. The odds of the hypotheses (the domain of the trier of fact) are dependent on all evidence and information, while the expert only has relevant information about the evidence within his field of expertise. An expert witness contributes to a trial by providing information about the probability of the evidence in the expert's domain (i.e. (dis)similarities in chromatograms for a gasoline comparison) under the two hypotheses. The ratio of these two probabilities is the likelihood ratio. The LR has values between 0 and infinity. Values smaller than 1 support the defense hypothesis (H_d), and values larger

than 1 support the prosecution hypothesis (H_p) [1-7,9]. A value of 1 represents neutral evidence. Larger LRs give stronger support for H_p and LRs closer to zero give stronger support for H_d.

A LR approach is new to the field of gasoline comparison. Previous studies of evaporated gasoline residue comparison approach the problem as a classification problem. In these studies, samples of several sources of gasoline are prepared and it is studied whether numerical techniques can group gasolines from the same source. There is an important difference between a classification approach and a LR approach. Classification methods make categorical decisions based on the comparison only, while a LR provides the evidential value of the comparison result. The latter allows for logical combination with other evidence and information, and thus allows the trier of fact to decide based on all information available, while the former does not.

1.2 Comparison of evaporated gasoline residues

Even though the usefulness of classification studies for forensic purposes is limited, they do provide relevant information on which features of the chromatogram to use for a LR approach. We will therefore briefly describe a number of classification studies.

Some early work has been done on discriminating gasolines from different sources by fluorescence spectroscopy [10,11], but for most of these approaches the chemical composition is identified by gas chromatography (GC) [12-17]. This is done for a number of gasoline samples for which the amount of evaporation is varied under laboratory conditions. A first approach, pioneered by Mann [14,15] was based on analyzing chromatographic peak area ratios of volatile compounds. The use of volatile compounds limited the method to gasolines that were no more than 50% evaporated. This method was extended by Barnes et al. [12] to the qualitative comparison of a selection of peak area ratios including peaks at longer retention times and samples up to 75% evaporation. Peak area ratios (4 to 6 depending on the amount of evaporation under consideration) were selected based on minimal variation within a gasoline sample for varying amount of evaporation, and good discrimination between gasoline samples.

A second approach combines GC with statistical methods in order to reduce the dimensionality of the chromatogram data. In this approach the data is described by fewer variables than peaks in the chromatogram, while capturing a considerable amount of variance in the data. Sandercock & Du Pasquier [17] used principal component analysis and linear discriminant analysis to discriminate samples up to 90% evaporation. They used the C_0 to C_2 naphtalene composition (11 late-eluting peaks in the chromatogram) as input for statistical analysis. In their study, 35 samples of gasoline (each evaporated to 0, 25, 50, 75, and 90%) were found to form 18 groups in a linear discriminant analysis; 11 of these contained a single gasoline, while the other 7 groups contained 2 to 6 gasolines.

Recently, Petraco et al. [18] studied the performance of a variety of multivariate statistical techniques in order to discriminate between 20 retained liquid

gasoline samples from fire investigations. They selected 15 peaks from the chromatogram as input for statistical analysis. The 15 peaks were selected based on their consistent vapor pressure up to 75% evaporation. In order to create groups, a number of replicated measurements were made per gasoline.

1.3 Scope of the present study

In the present study the goal is to obtain numerical values for LRs calculated for the comparison of evaporated gasoline residues to other evaporated gasoline residues and unevaporated gasolines. Three methods to calculate LRs are presented. The methods differ in the features used to discriminate between gasoline residues and in the statistical approach used to obtain LRs. Two of them build on the literature on classification of evaporated gasoline residues and the forensic statistical literature. The third comparison method has not been published previously in the literature on gasoline comparison.

The contribution of this paper is twofold. 1. Introduce the LR-approach to the field of forensic gasoline comparison. 2. Take a first step in the creation of computer-based methods in order to assign an evidential value to gasoline comparisons for forensic purposes, by accounting for evaporation.

2. Materials and methods

2.1 Hypotheses

In this work source level hypotheses are addressed [19]. Different-source gasoline samples are defined as either coming from different petrol stations or from the same petrol station that has been refilled in the meantime. Same source gasoline samples come from the same petrol station and refill. The definition of the same source and different source hypotheses in this way is in accordance with the experience of forensic experts on gasoline comparison at the NFI: gasoline at the tank of a petrol station is relatively stable in between refills.

In casework however, the relevant source of gasoline is at the person level. This is because different people who have collected gasoline from the same tank and refill are considered as different sources by the court. While gasolines from these persons should be considered as from a different source in casework, they are defined as from the same source in our dataset.

In order to obtain a more relevant dataset for casework, a survey of gasoline samples at people’s homes would be preferred. However, for the current purpose of assessing and comparing the performance of three different LR-methods the current dataset is appropriate. It is not the goal of the present work to obtain a LR-method to be used in casework, but to show the feasibility of LR-methods for the comparison of evaporated gasolines.

2.2 Data

Samples of gasoline were obtained by repeatedly collecting samples at 15 petrol stations in the region of The Hague in the Netherlands. At a petrol station gasoline was collected at one or two fuel pumps. When two fuel pumps were used, Euro 95 gasoline was collected from one fuel pump, and a high octane grade gasoline was collected from the other. The time between each sampling was one week or more and it was checked that in between two collections the fuel pumps had been refilled. Samples were collected from July to October 2009 (189) and in July 2010 (29) and September 2010 (29). The high octane grade fuels contained 7 types of gasoline. The 15 petrol stations were of 9 different brands: BP, Esso, Gulf, Shell, Tamoil, Tango, Texaco, Tinq, and Total. A total of 258 samples of gasoline were collected.

Evaporated samples (126) were prepared from 42 of the 258 unevaporated gasolines, by evaporation of unevaporated gasoline samples in the laboratory. Evaporation was continued until residuals contained 75%, 50% or 25% by weight. Half of the evaporated gasolines were Euro 95 gasoline, the other half were high octane grade gasolines.

The chemical composition of all of the samples was analyzed by gas-chromatography. An Agilent 6890N gas chromatograph equipped with an FID detector and an Agilent 7683 Series autosampler were used. FID detection was chosen for reasons of good reproducibility and robustness. In comparison to MS, the main advantage is that FID detection is more stable over time. The use of an FID detector allows for the comparison of a dataset of chromatograms that have been measured over a time span of a few years. The column consisted of 25 meter fused silica with an internal diameter of 0.20 mm and a methyl silicone stationary phase (Ultra 1) of 0.33 µm thickness. The initial temperature after injection was 50°C, and temperature was ramped with 2°C/min up until 160°C. Subsequently the temperature was increased to 250°C by 30°C/min in order to elute high-boiling contaminants.

An automated procedure was used to integrate peaks in the chromatograms in order to obtain the integrated area by compound. An automated integration procedure is included in the Chemstation software package (Rev. A.09.03 (1417)) used to obtain the chromatograms. The areas computed by the automated procedure are dependent on the position of the baseline and the presence of shoulders from co-eluting compounds. Some of the compounds eluted as resolved baseline separated peaks, but others had shoulders from co-eluting compounds or eluted in unresolved multi-component peak patterns. Consistency in the reconstruction of the baseline by the integration algorithm, when integrating repeated measurements, is critical. The more complex the peak pattern, the harder this task gets. 53 peaks were selected that showed consistency in the reconstruction of the baseline and minimal variation in repeated measurements. These were mainly baseline-separated peaks, or peaks with only small shoulders. Peak-areas were calculated for these compounds and these were retained for further analyses.

From this data, three datasets were created:

1. A training data set of evaporated gasolines was created from 22 of the 42 gasolines used for evaporation. These gasolines were each evaporated to approximately 25, 50, and 75% by weight, leading to 66 samples.
2. A validation dataset of evaporated gasolines was created from the remaining 20 gasolines used for evaporation. These gasolines were each evaporated to approximately 25, 50, and 75% by weight, leading to 60 samples.
3. The data from the 258 unevaporated gasolines were used as background dataset. For the multivariate method, the probability density over the feature space for the population of gasolines was estimated from this sample. Moreover, for different source comparisons, pairs of gasolines to be compared consisted of one from the background dataset and one from the training or validation dataset.

2.3 Selection of stable ratios

As suggested by the literature on classification studies for evaporated gasoline residues, selection of stable peak area ratios was adopted. A peak area ratio is the ratio of the areas of two peaks in a chromatogram. From the training dataset all peak area ratios were calculated and 13 peak area ratios were selected that showed little variation within a gasoline for the four evaporation levels, as compared with variation between gasolines of the background data. The procedure to select the peak area ratios follows below.

Area ratios were calculated for all peak pairs and a parameter F' (which is similar to an *F*-statistic) was calculated for each peak area ratio *R*. An *F*-statistic would have calculated the ratio between the variance of a peak area ratio within gasolines and between gasolines. Instead of calculating an *F*-statistic, the peak area ratios were standardized by the mean peak area ratio within a gasoline (averaged over evaporation levels). This was done for two reasons. 1. Peak area ratios vary substantially over gasolines, and it is desired to have a number which is independent of the size of the peak area ratio. 2. Random measurement error is proportional to the value of the peak area ratio, so that the division operation standardizes this error.

This resulted in the following statistic for a peak area ratio *R*,

$$F'(R) = \frac{1}{n(m-1)} \sqrt{\sum_i \sum_j \left(\frac{R_{ij}}{\overline{R}_i} - 1 \right)^2} \times \frac{\overline{R}}{sd_R}. \qquad \text{(Eq. 1)}$$

Here, *n* denotes the number of gasolines in the training dataset, and *m* denotes the number of evaporation levels, *ij* denotes gasoline *i* with evaporation level *j*. $\overline{R}_i$ denotes the mean peak area ratio of gasoline *i* over evaporation levels *j*. Whereas the term on the left is the (scaled) within standard deviation, the term on the right is the inverse of the (scaled) between standard deviation. sd_R is the standard deviation of peak area ratio area *R* over the background data. $\overline{R}$ is the mean of *R* over the

background data. Peak area ratios were selected to minimize F', while making sure that a peak occurred only once in the selected ratios. If a peak would occur more than once in the selected ratios, the discrimination of the peak area ratios would be less due to a dependency of the peak area ratios. Keeping F' below 0.05, 13 peak area ratios were retained.

Table 1 gives an overview of the compounds whose areas were used in the ratios. Note that for the numerator and denominator of the peak area ratios, peaks with similar retention indices (for a definition, see [20]) were selected by the automated procedure. This is explained by the relation between retention time and vapor pressure. Selection of compounds with similar vapor pressure (and thus retention time) will result in ratios that are relatively independent of the level of evaporation. The stability of the peak area ratios as a function of evaporation level may be assessed from

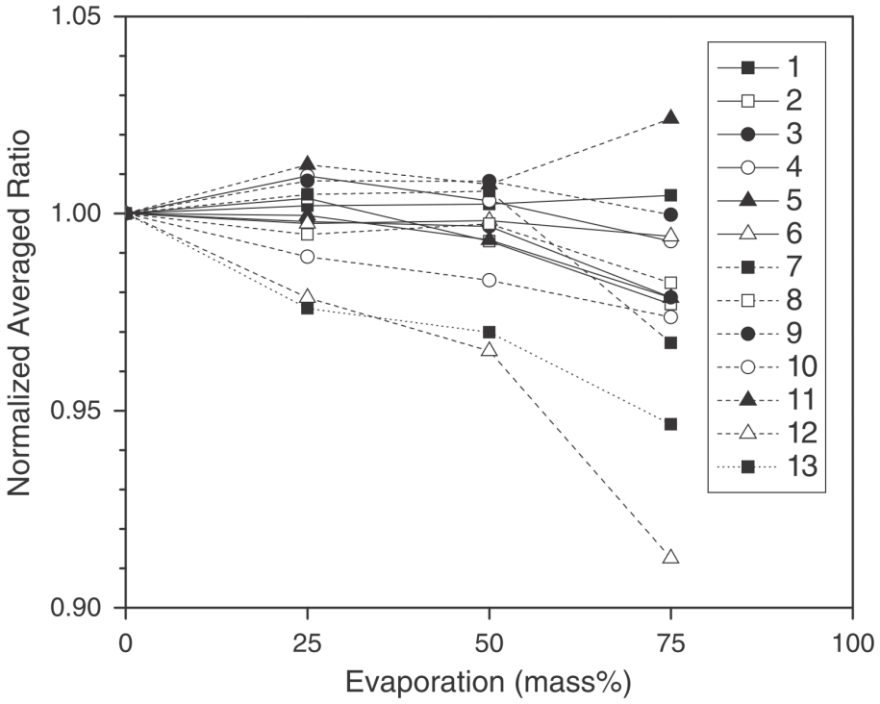

Fig. 1. It shows that as a function of evaporation level, all averaged normalized peak area ratios (first normalized to the ratio at 0% evaporation and subsequently averaged over gasolines) remained well within 10% of the value at 0% evaporation. Also note that most peak area ratios showed a slightly decreasing trend as a function of evaporation level. This may be explained by the fact that in all selected peak area ratios the numerator is a compound with a shorter retention time than the denominator. Since in general retention times are positively correlated with vapor pressure it may be expected that compounds with shorter retention times evaporate faster, yielding a decrease in the value of the ratios as a function of evaporation level. However, the peak area ratio selection procedure was optimized to select the peak area ratios that are most stable with respect to evaporation.

Note that the peak area ratios plotted in

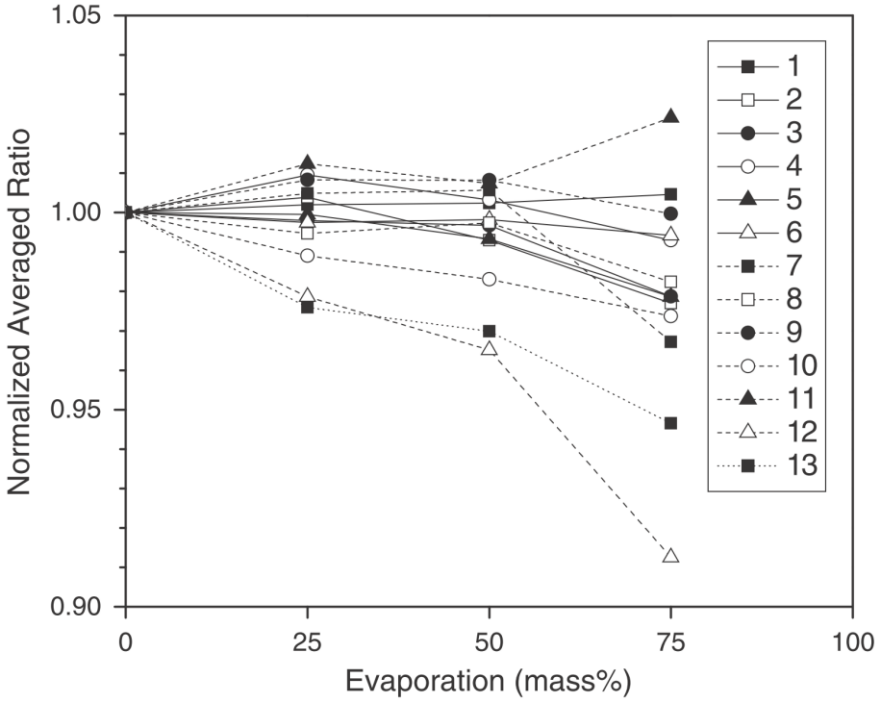

Fig. 1 are the peak area ratios selected based on the evaluation of Eq. 1. This equation evaluates peak area ratios not only on minimum variation within a gasoline, but also relative to the variation between gasolines. Peak area ratios that showed relatively large variation in

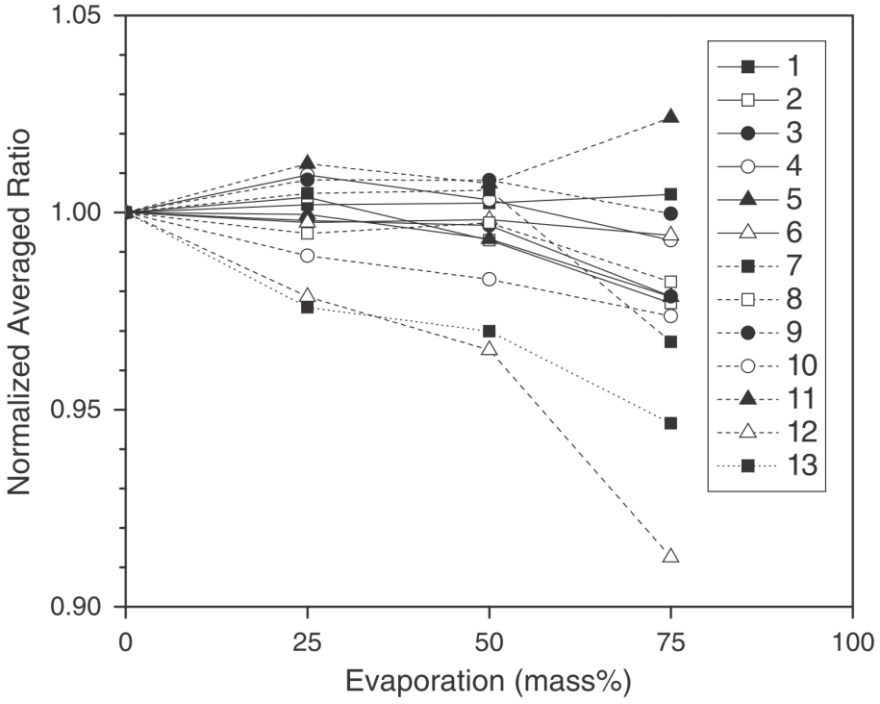

Fig. 1 also had large variation between gasolines.

2.4 LR formulas

The LR is defined as

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

Or for continuous data

$$LR(E) = \frac{f(E \mid H_p)}{f(E \mid H_d)}, \qquad \text{(Eq. 3)}$$

where *f* denotes a probability density.

The numerator of this equation is the probability density of observing the evidence E when H_p is true. The denominator of this equation is the probability density of observing the evidence E when H_d is true. In forensic LR-methods, distance (or similarity) measures between the features of the compared items are often used as the evidence. In that case, probability distributions of distances under H_p and H_d are obtained from a data set of distances. In the present paper, the dataset under H_p consists of data from gasoline pairs that have the same source. The dataset under H_d consists of data from gasoline pairs that have different sources. A LR is evaluated at distance d as,

$$LR(d) = \frac{f(d(\mathbf{x},\mathbf{y}) \mid H_p)}{f(d(\mathbf{x},\mathbf{y}) \mid H_d)}, \qquad \text{(Eq. 4)}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote the vector of features to be compared of gasoline X and Y and $d(\mathbf{x},\mathbf{y})$ is the distance function.

The distance function based on selected ratios, d_1

For the first distance function d_1 (based on the 13 selected ratios), when comparing an *x*% evaporated gasoline (denoted for peak area ratio *j* as $R_{jx\%}$) to an unevaporated gasoline (denoted for peak area ratio *j* as $R_{j0\%}$), d_1 is defined as

$$d_1 = \sqrt{\sum_{j=1}^{13} \left(\log\left(R_{jx\%} / R_{j0\%}\right)\right)^2}. \qquad \text{(Eq. 5)}$$

The following reasoning was applied to arrive at this distance function: For a same-source gasoline comparison, ratios $R_{jx\%}$ and $R_{j0\%}$ are approximately equal and division standardizes the random component of the measurement error. The use of the

square of the logarithm ensures that the distance function is indifferent to the order of comparison (i.e. using $R_{jx\%} / R_{j0\%}$ or $R_{j0\%} / R_{jx\%}$) and a log transformation makes the method more robust to outliers. Several other distance functions based on the 13 ratios were explored and the present one was found to discriminate best.

The distance function based on trends in vapor pressure, d_2

The second distance function d_2 was based on the observation that compounds that are more volatile generally have shorter retention times than less volatile compounds. This is due to a strong correlation between retention time and vapor pressure. Thus, the ratio of the area under a peak in a chromatogram of an evaporated gasoline to the area for a same-source unevaporated gasoline is smaller for peaks with shorter retention times (see

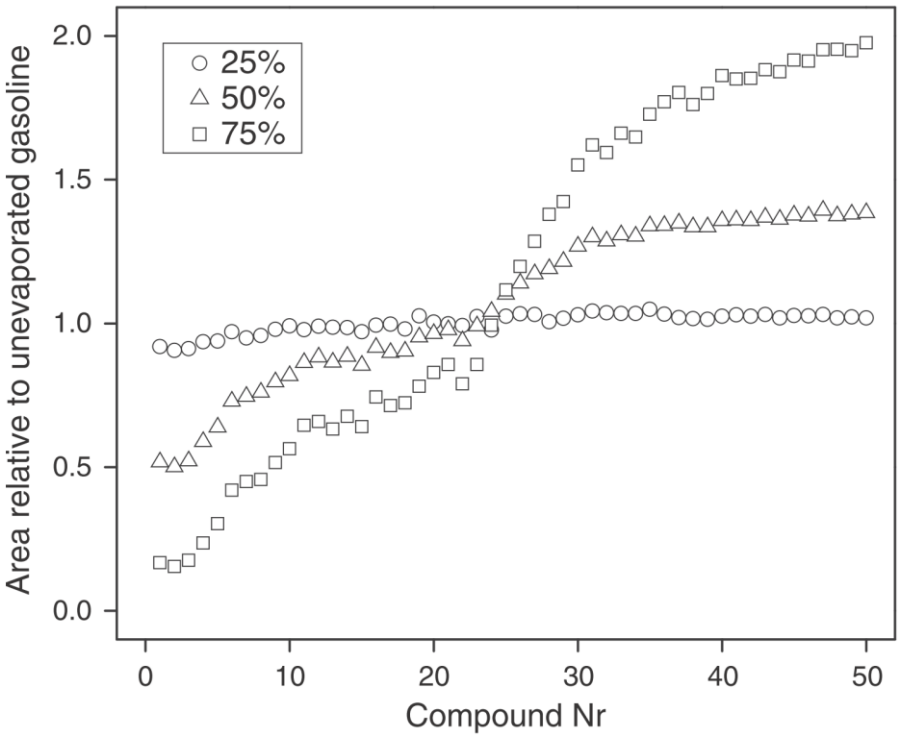

Fig. 2).

When comparing the chromatograms of two gasolines, if a ratio for a peak at shorter retention time is found to be larger than a ratio for a peak at later retention time, this provides evidence for the two gasolines coming from a different source. This observation motivates the following distance measure for peak *j* and peak *i* at shorter retention time ($RT_i < RT_j$).

$$d_{ij} = (y_{ij} - 1)^2 \text{ for } y_{ij} > 1, \text{ and} \qquad \text{(Eq. 6a)}$$

$$d_{ij} = 0 \text{ for } y_{ij} \leq 1, \qquad \text{(Eq. 6b)}$$

where $y_{ij} = \frac{a_{ix\%} / a_{i0\%}}{a_{jx\%} / a_{j0\%}}$. The peak areas *a* of peak *i* and peak *j* of a *x*% evaporated gasoline are denoted by $a_{ix\%}$ and $a_{jx\%}$ while the peak areas *a* of peak *i* and *j* of a 0% evaporated gasoline are denoted by $a_{i0\%}$ and $a_{j0\%}$.

The distance function d_2 is given by

$$d_2 = \sqrt{\frac{1}{n}\sum d_{ij}}, \qquad \text{(Eq. 7)}$$

where *n* is the number of y_{ij} that were larger than 1. Note that *n* may vary over gasoline comparisons. Its function is to suppress the contribution of many (and small) d_{ij} due to measurement error when a same-source gasoline comparison involves gasolines with approximately equal evaporation levels (for example when comparing the 0 and 25% evaporated gasolines of

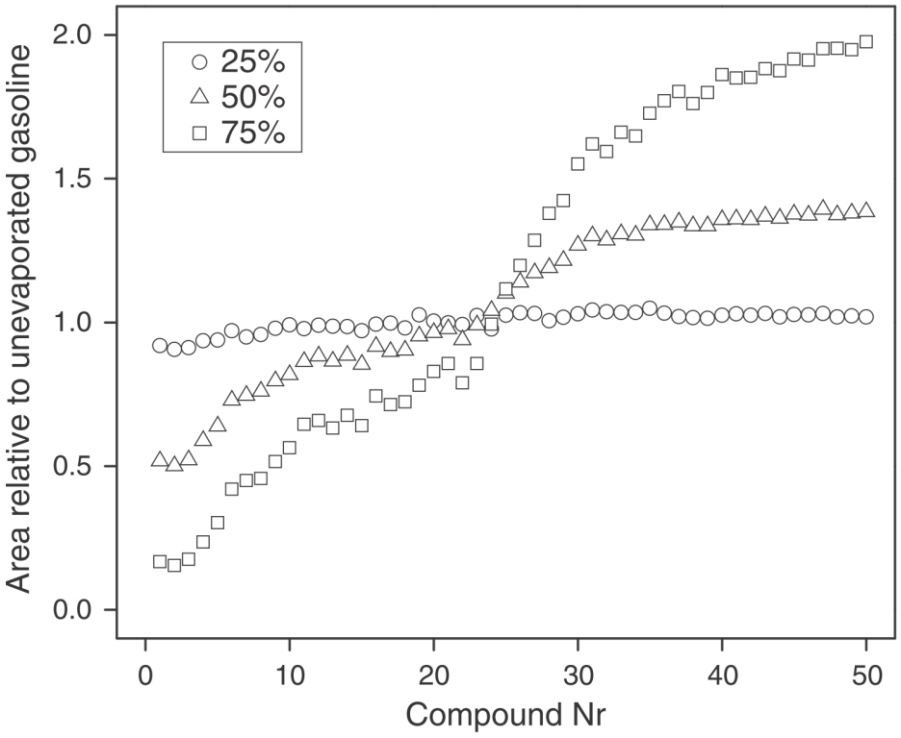

Fig. 2).

For both distance methods, kernel smoothing was performed on the distributions of distances using a Gaussian kernel. The criteria used to select the kernel bandwidth were (a) a good visual correspondence between the kernel density and the histogram in the region of overlap between the two distributions and (b) a resulting transformation function from score to LR that is monotonously decreasing. The values for the kernel bandwidth for d_1 are: $h = 0.020$ (same source) and $h = 0.026$

(different source). The values for the kernel bandwidth for d_2 are $h = 0.0057$ (same source) and $h = 0.0060$ (different source).

Multivariate distribution method

The third approach for the calculation of LRs was to model the probability distributions of characteristics directly and calculate the probability density for joint occurrence of characteristics $\mathbf{x}$ and $\mathbf{y}$ given the two hypotheses,

$$LR(\mathbf{x},\mathbf{y}) = \frac{f(\mathbf{x},\mathbf{y} \mid H_p)}{f(\mathbf{x},\mathbf{y} \mid H_d)}. \qquad \text{(Eq. 8)}$$

In contrast to the probability distributions based on distance methods, these probability distributions are multivariate in nature. Therefore, we call this approach the multivariate distribution method [4]. Following this approach, a LR based on the ratio of the two probabilities of all selected variables (in this case two times 13 peak area ratios) occurring under the two competing hypotheses was used. The probabilities of the peak area ratios within the same source were assumed to follow a multivariate Gaussian distribution. For this LR method a covariance matrix is estimated to model the within source variation. In order to estimate stable within source and between source probability densities, dimension reduction was performed by principal components analysis (PCA). PCA was performed on the correlation matrix of the background data. The PCs with eigenvalues larger than one were retained in the subsequent LR calculations. This amounts to 3 PCs explaining 79% of the variance.

The covariance matrix was estimated by a mean within covariance S,

$$S = \frac{\sum_{i=1,\, j=1}^{i=19 \text{ or } 20, j=4} (\mathbf{y}_{ij} - \mathbf{y}_i)(\mathbf{y}_{ij} - \mathbf{y}_i)^T}{n(m-1)}, \qquad \text{(Eq. 9)}$$

where n denotes the total number of gasolines, and m the total number of evaporation levels. The vectors $\mathbf{y}_{ij}$ contain the scores for the 3 PCs for gasoline i with evaporation level j.

Since the selection of the ratios was done by minimizing variances of the training data, the training data cannot be used to reliably estimate the within covariance matrix. Therefore the validation data were also (apart from calculating LRs) used to estimate the mean within source covariance matrix and the following validation protocol was used.

For same-source comparisons a leave-one-out cross-validation scheme was used. For a particular gasoline pair a mean covariance matrix was estimated from the other 19 groups of gasolines in the validation dataset. For different source comparisons the mean covariance matrix was estimated using all 20 groups of gasolines from the validation dataset. This is appropriate since for different-source LR

calculations the background database was used, which is not involved in estimation of the mean within-source covariance matrix.

For estimation of the probability distribution of different-source gasolines an empirical distribution was used, which was smoothed with multivariate kernel density estimation [21]. LR formulas may be found in the appendix.

3. Results

For the LRs based on distance functions, distances were obtained by comparisons of gasolines from the training dataset to the background dataset. For same-source comparisons, a distance was computed for a comparison of each 25, 50, or 75% evaporated gasoline from the training data to its unevaporated counterpart from the background data. This amounts to 66 distances for same-source comparisons. Distances for different-source comparisons were computed for comparisons of each 25, 50, or 75% evaporated gasoline from the training data to each different gasoline from the background data. This amounts to 16962 ($22 \times 3 \times 257$) distances for different-source comparisons.

Fig. 3a and 3b show histograms of distributions of same-source and different-source gasoline comparisons for both distance functions. The *y*-scale of the graphs is optimized to show the probability density of different source comparison distances. The scale of the insets is optimized to show the histograms of the same-source comparisons. The lines show the kernel density approximations, which follow the empirical distributions. Both distance methods show little overlap between different and same-source comparisons, reflecting good discrimination properties. Most mass of the density of same-source comparisons is concentrated at small distances, whereas the density of different-source comparisons shows a much larger spread. For d_1 some same-source comparisons have larger distances, leading to a wider profile as compared to the same-source distribution of d_2.

LRs were calculated for the validation data. For same-source gasoline comparisons, each 25, 50, and 75% evaporated gasoline from this dataset was compared to its unevaporated counterpart from the background dataset. This amounts to 60 LRs for same-source comparisons. It would be convenient to have (much) more data available, but the work of preparing samples is quite substantial. For different-source comparisons, each 25, 50, or 75% evaporated gasoline from the validation dataset was compared with each different gasoline from the background dataset. This amounts to 15,420 LRs for different-source comparisons.

Multivariate LRs were also calculated for same-source and different-source gasoline comparisons. In order to reliably estimate the covariance matrices, PCA was used as a dimension reduction technique. Fig. 4 shows the PC scores for the validation data (black) and the background dataset (open circles). A clustering is observed for the same type of gasoline, but at different evaporation levels. The clustering is much stronger than the variation in the background population, indicating good

discrimination properties of the multivariate LR method. The distribution of the validation data is similar to the distribution of the background data, reflecting that the validation data is a sample from the background data. Note that for the validation data evaporated gasolines were used and that those ratios were selected with the goal to suppress the effect of evaporation.

Fig. 5 compares the LRs for same-source comparisons of the three LR methods by showing a plot of the proportion of same-source comparisons that have $\log_{10}$ LRs larger than the value on the x-axis. For the multivariate method, the majority of comparisons yielded LRs in the range 10^2 to 10^5. The median value was about 9×10^3. The two distance methods gave comparable results and yielded LRs in the range of 10^2 to 10^4 (median LR for the distance function for trends in vapor pressure (d_2) was 3.1×10^3, and median LR for the distance function for 13 ratios (d_1) was 2.7×10^3).

The fraction of misleading evidence for same-source comparisons (yielding LR < 1) varied for the three methods, as may be seen from Fig 5. The multivariate method yielded most misleading evidence (13.3%). Distance method d_2 had three occurrences of misleading evidence (5.0%), method d_1 yielded one occurrence of misleading evidence (1.7%).

The large proportion of misleading evidence for d_2 is a surprising result. It is caused either by the presence of a few outliers in the same-source validation data, or conversely by the lack of such data in the tail region of the same-source training data. Since no experimental changes were made in the collection process of the validation and the training data, and the training data was not used for variable selection for d_2, the only reasons to account for this are the sampling process or the selection of the kernel bandwidths.

The LR-values for the multivariate method were more strongly misleading than those for the distance methods. The relatively large rate of misleading evidence for same-source comparisons for the multivariate distribution method may be explained by the modeling assumptions intrinsic to this method. Distributions with identical means are assumed for same-source gasolines. This may not be the case when comparing evaporated gasolines to unevaporated gasolines. Although the ratios were selected for their stability with respect to evaporation, small deviations as a function of evaporation level were observed. This may lead to deviations in means large enough to yield very strongly misleading evidence.

An inspection of the same-source comparisons that resulted in misleading evidence showed that the three LR methods were rather consistent. For one same-source comparison, all three methods yielded misleading evidence. For the same-source comparisons for which the multivariate method yielded misleading evidence, the other two methods yielded LRs which were among the smallest in their series.

An overview of misleading evidence for different-source comparisons (LR > 1) is given in Table 2. For all three methods the overall fraction of misleading evidence was smaller than 1%. It was largest for the multivariate LR method, whereas it was smallest for the distance function based on trends in vapor pressure (d_2).

For d_1 the rate of misleading evidence for the different comparisons is larger than for d_2. This difference may be explained by the larger number of peak pairs involved in the second distance function. For this method, all combinations of two peaks may contribute to the distance. This amounts to 1378 peak pairs (although the contributions of the peak pairs are not independent), as compared with 13 peak pairs involved in the ratio methods. Since more information is likely to be contained in the larger amount of peak pairs used for d_2, this may explain the better performance of the LR method for different comparisons for d_2. From this perspective, it is reassuring that d_1 performed so well, with only 13 ratios included. The inclusion criteria for selection of ratios were chosen to find the optimal performing ratios. Apparently, this selection procedure worked rather well.

For both distance methods the majority of misleading evidence for different-source comparisons had LRs between 1 and 100 (71% for d_2 and 80% for d_1). For the multivariate distribution method, such LRs were roughly evenly distributed on a logarithmic scale between 1 and 10^5.

Inspection of the different-source comparisons for which strongly misleading evidence was obtained revealed that most of these gasoline pairs were collected within 2 weeks at different pumps. The occurrence of strong similarities between a few gasolines in the dataset is not surprising since some of the petrol stations used for sample collection were serviced by the same tanker. Therefore, strong similarities between some samples are expected.

3.1 Assessment of calibration

A LR is defined as the ratio of two probabilities (or probability densities). A LR method gives calibrated LRs if the probability distributions are modeled accurately. For the two distance functions the probability distributions were estimated from empirical data using kernel density estimation. For the multivariate LR-method a statistical model was assumed for the within distribution (a multivariate normal distribution) and the between distribution was estimated from empirical data by kernel density estimation.

Calibration of assigned probabilities may be assessed by comparing a calculated probability by the relative frequency of this probability in a test data set. If the relative frequency is equal to the estimated probability, the probability estimate is well calibrated. A graphical measure of such an assessment is called an empirical calibration plot [22], where estimated probabilities are plotted as a function of observed relative frequencies.

In order to assess the calibration of the LRs, the following relation [23] is helpful, which is only true for perfect calibration:

$$LR = \frac{p\left(LR \mid H_p\right)}{p\left(LR \mid H_d\right)}, \qquad \text{(Eq. 10)}$$

This equation is analogous to the definition of the LR in Eq. 2, the difference being that the LR is also interpreted as the evidence in the right hand side of the equation. Since the LR contains all important information about the evidence relevant for the discrimination between H_p and H_d, Eq. 10 holds. This implies that when a method claims to output LRs (denoted as method-LRs), then generating a series of method-LRs under H_p and H_d, calibrating them and plotting the calibrated LRs as a function of method-LRs yields a line $y = x$. Given that the latter calibrating process has been performed well, deviations from this line denote that method-LRs are not properly calibrated. This property will be exploited below.

In order to extrapolate a LR-method beyond its training data, the property of calibration should hold for the LRs of the validation data. In order to measure calibration for the validation data, LRs for this data may be calculated and calibration may be assessed by assessing whether Eq. 10 holds for this set of LRs.

As an empirical calibrating method for the method-LRs for the validation data of the three different methods under study, the PAV algorithm (originally developed to find posterior probabilities in [24], adopted to find LRs in [25]) was used. This algorithm transforms method-LRs into calibrated LRs for the present dataset under the constraint of monotonicity. Monotonicity means that larger method-LRs are associated with larger calibrated LRs. The algorithm uses ordered method-LRs (all H_p and H_d method-LRs are used in one sequence) as input and bunches them so that the ratio of H_p to H_d method-LRs in one bunch results in a stepwise increasing number over the bunches. These ratios are subsequently set equal to posterior odds and converted to LRs using Bayes' rule and the fact that the prior odds are defined by the ratio of the number of data under H_p and H_d. These prior odds are determined by the conditions of the experiment. Since the posterior odds obtained by the PAV algorithm are equal to the LR × prior odds, division of the PAV result by the prior odds gives the desired result. When method-LRs are invariant under this transformation, calibration for the method-LRs was already perfect and a plot of transformed LRs versus method-LRs should yield a line close to the line $y = x$.

Figure 6 shows the PAV transformation results for the three different methods. The PAV transform is the stepwise increasing line in the plot. As a visual aid, the line $y = x$ is plotted in the Figures. For the first distance method the PAV transform is plotted in Fig. 6a. It shows that the line $y = x$ is followed quite closely in the region from log method-LRs 1 to log method-LRs 4. Therefore, in this region calibration is good. Log method-LRs of -2 to 1 are transformed to a log PAV calibrated LR of 0.5, which shows that calibration is bad. For log method-LRs smaller than -2 and larger than 4 calibrated LRs are not supported due to the lack of data under H_p on the left hand side and under H_d on the right hand side.

In Figure 6b the PAV transform for the LR method based on the second distance method is shown. It shows the same trend as for the first method, except that the region of small method-LRs for which calibration is bad is extended. The region now goes from log method-LR -7 to 1. It is not surprising that the limits on the left hand side for the two methods (-2 and -7) coincide with the smallest method-LRs obtained

under H_p in Fig. 5. These are the smallest method-LRs that can be transformed by the PAV algorithm.

Figure 6c shows the PAV transform for the multivariate method. Note that small method-LRs have very bad calibration, log method-LRs of -15 being transformed to 0.5. Also note that calibration is off for the complete range of log method-LRs from -15 to 5. Log method-LRs from 2 to 5 are transformed to values roughly two orders of magnitude smaller. That calibration is further off for the multivariate distribution method as compared with the distance methods may be explained by the use of a statistical model for the within distribution (a multivariate normal model) as compared with the empirical models that were used for the methods based on distance functions. For the multivariate model, the empirical distributions may not follow a multivariate normal distribution. Moreover, for the parameter estimation a mean within covariance matrix was used, which may deviate from the (hypothetical) covariance matrix for individual gasolines.

3.2 Measuring performance by an information theoretical approach

In order to further investigate the performance of the three methods, a measure based on information theory was used. This measure is the Empirical Cross-Entropy (ECE), which has been proposed as a measure to analyze the performance of LR-methods (see e.g. [26]). ECE has been applied to characterize LR-methods in [22,26-30]. A detailed description of how to use ECE and the PAV algorithm to measure the performance of other physicochemical data may be found in [31]. The ECE measures the accuracy of probability statements by applying a logarithmic scoring rule. The further the probability is from predicting the ground truth, the larger the cost, with no upper bound. The lower bound is defined by the cost for perfect predictions (e.g. probability assignments of 1 for H_p as the ground truth and 0 for H_d as the ground truth). In this case the ECE yields 0.

As was noted above, the ECE measures the accuracy of probability statements. In order to apply the ECE to LR-methods, a specification of the prior probability is mandatory. Because we don't know the prior probability this is solved by calculating the ECE for the LR-method over a range of prior probabilities and plotting the ECE as a function of prior probability. The advantage of this approach is that it is in accordance with the roles of the actors in court, where the role of the expert is to communicate the likelihood ratio of the evidence and the prior probability is the province of the other actors.

An advantage of the logarithmic scoring rule is that it is a strictly proper scoring rule. This means that in an ECE plot, the performance can be separated in a part measuring the discrimination and a part measuring the calibration, given that the calibrating procedure used preserves the discrimination properties. The discrimination properties are preserved because a monotonous transform for the LR to calibrated-LR is expected.

LRs are said to be well calibrated when both probabilities (in the numerator and the denominator) are assigned accurately. When one or both of these probabilities

do not correspond to the proportions in the sample, this will result in a larger ECE. In order to measure the calibration loss, the LRs of the validation data may be (re)calibrated using their frequency of occurrence under H_p and H_d. The PAV-algorithm incorporated in the ECE analysis is a calibration method. The ECE score after such a calibration is a measure of the discrimination.

Apart from a curve showing the ECE of the LR-method and a curve showing the ECE of a method calibrated on the validation data, the ECE-plot also shows a curve for uninformative evidence, where the LR always equals one. This may be used as a benchmark for minimal performance of the LR-method. Note that for LR-methods with bad calibration, the LRs produced may be worse than uninformative LRs.

In Fig. 7 the ECE plots are shown for the three methods. In (a) the ECE plot for the first distance method (based on the selected ratios) is shown. The solid curve represents the ECE of the LR-method, while the dotted line represents the ECE of LRs equal to one. The dashed curve represents the ECE of the method where the calibration is optimized for the validation data by application of the PAV algorithm. The ECE of the LRs produced by the LR-method (solid curve) is much smaller than the ECE of uninformative LRs. After PAV calibration, the performance is improved further, showing that the KDE fitted on the training data is not optimal for the validation data.

ECE curves for distance method 2 (based on trends in evaporation) are shown in Fig. 7b. Note that the performance of this method strongly depends on the prior odds. For negative log prior odds, performance is good, while for positive log prior odds a large ECE is obtained, even much larger than for uninformative LRs. This large outcome is explained by the presence of three misleading LRs that are very small while H_p is true, which are also visible in Fig. 5. For positive prior log odds these misleading LRs give a large contribution to the ECE.

After PAV transformation, the ECE improves drastically. The ECE is small in the whole plotting range, and comparable to the ECE of the PAV calibrated LRs for the first distance method in Fig. 7a. Therefore, the discrimination of the distance methods is comparable.

ECE curves for the multivariate distribution method are shown in Figure 7c. The same trend is visible as in Fig. 7b, but even more pronounced. This is due to four even more strongly misleading LR values under H_p. Fig. 5 shows that the four same-source comparisons yield LR values smaller than 10^{-10}. Three of these originate from comparisons between 75% evaporated gasolines and their unevaporated counterparts, while the other one is from a 50% evaporated gasoline (and its unevaporated counterpart). Possibly, the strongly misleading LRs may be explained by the larger amount of evaporation for these samples. This explanation is supported by Fig. 1 where it is shown that deviations of the normalized average ratio are strongest for 75% evaporation. The PCA factor scores used in the multivariate normal approach are composed of these ratios. The multivariate normal distribution is dependent on the difference in PCA scores, and larger differences lead to smaller probabilities under H_p and therefore smaller LRs. Note that after calibration of the LRs of the validation data

by PAV the ECE improves dramatically, showing the potential of the method. However, the PAV calibrated curve is not as low as for the two distance methods, showing that the two distance methods have superior discrimination.

3.3 Comparison of gasoline residues from gasolines that are both evaporated

Finally, it was also investigated whether application of the three methods could be extended to the comparison of two evaporated gasolines. For this purpose LRs were calculated for same and different-source comparisons of 25% evaporated gasolines to 50 and 75% evaporated gasolines. The validation dataset was used for this purpose. The use of this dataset is equivalent to 40 same-source and 760 different-source comparisons.

Results were in accordance with the results on evaporated-to-unevaporated-gasoline LR-results described above. Median values of the LRs for same-source comparisons were 3×10^3 for the two distance methods and 1.0×10^4 for the multivariate method, which were typical values found for same-source gasoline comparisons of evaporated to unevaporated gasolines. The amount of misleading evidence for same-source comparisons was 5.0% for method d_2 and the multivariate distribution method while it was 2.5% for method d_1. The rate of misleading evidence for different-source comparisons was 0.52% for method d_1, 0.66% for method d_2, and 0.13% for the multivariate distribution method.

ECE plots for the three methods are shown in Fig. 8. They show the same trends as for the comparison of evaporated to unevaporated gasolines. Again, method d_1 (based on the selected ratios) performs better than the other two LR-methods. The solid line of Fig 8a. is markedly lower than the line for LR =1 always, while for the other two methods in a major part of the graph the solid line is above the line for uninformative evidence in a major part of the graph. For the three methods the discrimination (dashed line) is comparable, which is in line with the reported similar median values of the LR under H_p and the rates of misleading evidence.

4. Discussion

The results show the potential of three LR methods to calculate LRs for evaporated gasoline comparisons, showing good discrimination behavior when an unevaporated gasoline is compared with a gasoline evaporated under laboratory conditions. The methods can also be applied to the comparison of residues of two evaporated gasolines. Two distance function based methods were developed and a third method modeling of the multivariate distribution of features was applied. This is a first step towards the development of automated methods in order to calculate LRs for gasoline residues found in casework. Following steps are to incorporate the effect of the matrix material on which the gasoline sample was found and microbial degradation, and to collect data in accordance with hypotheses used in casework.

Another benefit of collecting new data is the complete disentanglement of validation and training datasets. In the current experimental setup, disentanglement was pursued but was not perfect. For pragmatic reasons, the background dataset of unevaporated gasolines was used for multiple purposes. It was used in the selection of stable peak area ratios, for calculating distances for different-source comparisons and for the different-source distribution of the multivariate method. This may lead to overoptimistic results, since the background data was used in the training phase as well as in the validation phase. However, this bias in the results is expected to be small. Since independent sets of evaporated gasolines were used for training and validation, and these are paired with the background dataset to model and calculate LRs for different source comparisons, a possible bias in the outcome of the different-source comparisons is expected to be small.

A validation dataset was used to assess whether the developed methods generalize over gasolines. It was observed that the distance methods generalize better than the multivariate method, in terms of calibration. The PAV calibration results showed that the multivariate LRs were miscalibrated over the entire range, while the distance methods showed miscalibration for LRs smaller than 1. In terms of discrimination the distance methods outperformed the multivariate method.

In other fields (e.g. chromatographic comparison of MDMA profiles [4]) it has been observed that a multivariate distribution method performs superior to distance methods. The multivariate distribution method contains all information about the distributions while distance measures reduce the information. For gasoline comparisons, the discrimination of the multivariate distribution method was worse. This may be explained by the fact that the distance methods were designed to deal with the effect of evaporation while the multivariate distribution method was not. The multivariate distribution method assumes multivariate normal distributions with identical means for same-source gasoline comparisons. It has been observed that this assumption does not hold for evaporated gasolines, since the values of the selected ratios showed a trend as a function of evaporation (although they were selected to minimize this effect). This means that values of features of evaporated gasolines vary as a function of evaporation level, and are not identical when comparing gasolines at different evaporation levels. This may explain the relatively large rate of misleading evidence under H_p, and the lack of calibration of the multivariate distribution method.

The methods presented leave room for improvement. One example is the addition of a different number of ratios in ratio selection. Using more ratios may lead to better discrimination, but worse calibration. This is because better discriminating methods should yield larger LRs and more data is required to calibrate larger LRs. Using less ratios may lead to the reverse effect (better calibration at the cost of discrimination). In the present study the number of ratios was determined independent of these properties, by keeping F' below 0.05.

Another example is a refinement of the method based on trends in vapor pressure. The retention time is strongly correlated with vapor pressure, but the correspondence is not one-to-one. If all compounds are identified and vapor pressures

are known, the order of compounds could be based on vapor pressure directly instead of retention time as an indirect measure.

Further, it is noted that there is an intrinsic difference between the methods based on ratio selection and the method based on trends in vapor pressure. The former use symmetric functions whereas the latter uses an asymmetric distance function. This means that for the latter method the order of comparison is of importance: peak areas of the strongest evaporated gasoline are scaled by peak areas of the least evaporated gasoline. Thus, one has to determine which of the two gasoline residues under comparison is evaporated furthest in order to apply the LR method based on trends in vapor pressure. In a laboratory setting this knowledge is obvious, but this information may not be readily available in a case comparison, for example when gasoline residue found at the bottom of a jerry can at the suspect is also strongly evaporated.

Furthermore, a remark needs to be made about the evaporation levels involved. The methods presented in this work are currently limited to 75% evaporated samples. The question remains whether this maximum evaporation level suffices for casework. It would be interesting to be able to know the level of evaporation for gasoline residues in casework. To our knowledge, little work has been published in the literature on the subject of estimation of evaporation levels of gasolines. Hirz & Rizzi report accurate simulation of chromatogram data of evaporated gasolines up to 30% evaporation [32], and this may be used as a starting point to study the level of evaporation of gasoline residues. Extension of this approach to levels of evaporation higher than 30% has - to our knowledge - not been reported.

It is possible to extend the present analysis by, for example, inclusion of 90% evaporated gasoline samples. This was done for the distance method based on trends in vapor pressure. It was found that LRs for same source comparisons became small ($1 < LR < 50$). Also, for the ratio selection methods, it was observed that inclusion of 90% gasoline samples resulted in the exclusion of the first half of the compounds (with small retention times) in the ratios selected. This was interpreted as an undesired feature, since it reduces the discrimination properties of the gasoline comparison, and therefore it was decided to retain gasolines up to 75% evaporation. Future work on comparisons involving gasolines spiked on matrices will be more in accordance with casework. Matrices spiked with gasolines are left to evaporate for a time representative of casework. Validation of the methods on data of gasolines on matrices may provide insight in whether this maximum level of evaporation is sufficient.

5. Conclusion

To our knowledge, this study is the first one on the calculation of numerical LRs for evaporated gasoline comparisons. In this paper three LR methods were developed for a forensic comparison of an evaporated to a reference gasoline. The methods were validated for gasolines evaporated under laboratory conditions up to 75% evaporation. The three methods behaved well in terms of discrimination and rates of misleading evidence. The multivariate distribution method underperformed in comparison to the

two methods based on distance functions, possibly due to the fact that the assumption of identical means for same source gasoline comparisons does not hold. In terms of calibration, the multivariate method performed worst, giving miscalibrated LRs over the entire range, while the distance method based on ratio selection showed the best calibration properties.

In conclusion, it is possible to design LR-methods that discriminate well between same-source gasolines and different-source gasolines when gasolines have been partially evaporated. Future work will be aimed at extending the methods for application in forensic casework.

Appendix

$$\mathrm{LR} = \frac{f\left(\mathbf{y} \mid \mathbf{x}, H_p\right)}{f\left(\mathbf{y} \mid H_d\right)} = m \frac{\left|\mathbf{U}_{hn}\right|^{-1/2} \sum_{i=1}^{m} \exp\left\{-\tfrac{1}{2}\left(\bar{\mathbf{x}} - \bar{\mathbf{z}}_i\right)^t \mathbf{U}_{hx}^{-1}\left(\bar{\mathbf{x}} - \bar{\mathbf{z}}_i\right)\right\} \exp\left\{-\tfrac{1}{2}\left(\bar{\mathbf{y}} - \boldsymbol{\mu}_{hi}\right)^t \mathbf{U}_{hn}^{-1}\left(\bar{\mathbf{y}} - \boldsymbol{\mu}_{hi}\right)\right\}}{\left|\mathbf{U}_{h0}\right|^{-1/2} \sum_{i=1}^{m} \exp\left\{-\tfrac{1}{2}\left(\bar{\mathbf{x}} - \bar{\mathbf{z}}_i\right)^t \mathbf{U}_{hx}^{-1}\left(\bar{\mathbf{x}} - \bar{\mathbf{z}}_i\right)\right\} \sum_{i=1}^{m} \exp\left\{-\tfrac{1}{2}\left(\bar{\mathbf{y}} - \bar{\mathbf{z}}_i\right)^t \mathbf{U}_{h0}^{-1}\left(\bar{\mathbf{y}} - \bar{\mathbf{z}}_i\right)\right\}}$$

with $\mathbf{U}_{hx} = h^2\mathbf{T}_0 + n_x^{-1}S_x$, $\mathbf{U}_{h0} = h^2\mathbf{T}_0 + n_y^{-1}S_y$ and $\mathbf{U}_{hn} = \mathbf{T}_{hn} + n_y^{-1}S_y$ and

$\boldsymbol{\mu}_{hi} = h^2\mathbf{T}_0\left(h^2\mathbf{T}_0 + n_x^{-1}S_x\right)^{-1}\bar{\mathbf{x}} + n_x^{-1}S_x\left(h^2\mathbf{T}_0 + n_x^{-1}S_x\right)^{-1}\bar{\mathbf{z}}_i$ and

$\mathbf{T}_{hn} = h^2\mathbf{T}_0 - h^2\mathbf{T}_0\left(h^2\mathbf{T}_0 + n_x^{-1}S_x\right)^{-1}h^2\mathbf{T}_0$.

The vectors $\mathbf{x}$ and $\mathbf{y}$ denote features of gasolines X and Y. The vector $\mathbf{z}_i$ denotes the features of a gasoline *i* from a background sample Z. Matrix $\mathbf{T}_0$ denotes the covariance of the background sample, while in this work $S_x = S_y$ is the mean within covariance as defined in Eq. 9. The number of repeated measurements of X and Y are denoted by n_x and n_y (which are both 1 in this case) and *m* is the number of gasolines in the background sample. The parameter *h* is the optimal kernel bandwidth as defined in [21]. For a derivation of the formula, see [33] and references therein.

References

[1] C. G. G. Aitken and D. Lucy, Evaluation of trace evidence in the form of multivariate data, Applied Statistics 53, (2004) 109.

[2] C. G. G. Aitken and F. Taroni, in Statistics and the evaluation of evidence for forensic scientists, (John Wiley & Sons, Chichester, UK, 2004).

[3] C. G. G. Aitken, G. Zadora, and D. Lucy, A two-level model for evidence evaluation, Journal of Forensic Sciences 52, (2007) 412.

[4] A. Bolck et al., Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons, Forensic Science International 191, (2009) 42.

[5] C. Champod, Overview and meaning of identification, in: Edited by Siegel.J. Encyclopedia of Forensic Sciences,Academic Press, 2000).

[6] I. W. Evett, Towards a uniform framework for reporting opinions in forensic science casework, Science & Justice 38, (1998) 198.

[7] B. Robertson and G. A. Vignaux, in Interpreting evidence; Evaluating forensic science in the courtroom, (John Wiley & Sons, Chichester, UK, 1995).

[8] M. Trimpe and C. Chasteen, Comparing gasoline samples in the forensic laboratory, Fire & Arson Investigator (2002) 28.

[9] R. M. Royall, in Statistical Evidence: a likelihood paradigm, First ed. ed., (Chapman & Hall, London, New York, 1997).

[10] J. Alexander, G. Mashak, N. Kapitan, and J. A. Siegel, Fluoresence of petroleum products. II. Three-dimensional fluoresence plots of gasolines, Journal of Forensic Sciences 32, (1987) 72.

[11] L. M. Sheff and J. A. Siegel, Fluorescence of petroleum products. V. Three-dimensional fluorescence spectroscopy and capillary gas chromatography of neat and evaporated gasoline samples, Journal of Forensic Sciences 39, (1994) 1201.

[12] A. T. Barnes, J. A. Dolan, R. J. Kuk, and J. A. Siegel, Comparison of gasolines using gas chromatography-mass spectrometry and target ion response, Journal of Forensic Sciences 49, (2004) 1018.

[13] R. Hirz, Gasoline brand identification and individualization of gasoline lots, Journal of the Forensic Science Society 29, (1989) 91.

[14] D. C. Mann, Comparison of automotive gasolines using capillary gas-chromatography .1. Comparison methodology, Journal of Forensic Sciences 32, (1987) 606.

[15] D. C. Mann, Comparison of automotive gasolines using capillary gas-chromatography .2. Limitations of automotive gasoline comparisons in casework, Journal of Forensic Sciences 32, (1987) 616.

[16] T. L. Potter, Fingerprinting petroleum products: unleaded gasoline, in: Petroleum contaminated soils, vol. 2,Lewis publishers, Chelsea, MI, (1992).

[17] P. M. L. Sandercock and E. Du Pasquier, Chemical fingerprinting of gasoline - 2. Comparison of unevaporated and evaporated automotive gasoline samples, Forensic Science International 140, (2004) 43.

[18] N. D. K. Petraco, M. Gil, P. A. Pizzola, and T. A. Kubic, Statistical discrimination of liquid gasoline samples from casework, Journal of Forensic Sciences 53, (2008) 1092.

[19] R. Cook et al., A hierarchy in propositions: deciding which level te address in casework, Science & Justice 38, (1998) 231.

[20] H.J. Hubschmann, Handbook of GC/MS: Fundamentals and Applications, 2nd ed. Wiley-VCH Verlag GmbH, Weinheim, 2008.

[21] B. W. Silverman, in Density estimation for statistics and data analysis, first ed. ed., (Chapman and Hall, London, UK, 1986).

[22] D. Ramos and J. Gonzales-Rodriguez, Reliable support: Measuring calibration of likelihood ratios, Forensic Science International 230, (2013) 156.

[23] I. J. Good, Weight of Evidence: A Brief Survey, Bayesian Statistics 2 (1985) 249.

[24] R. E. Barlow, D. Bartholomew, J. M. Bremner, and H. D. Brunk, in Statistical inference under order restrictions; theory and application of isotonic regression, (Wiley, New York, 1972).

[25] N. Brummer and J. du Preez, Application-independent evaluation of speaker detection, Computer Speech and Language 20, (2006) 230.

[26] D. Ramos et al., Information-theoretical comparison of likelihood ratio methods of forensic evidence evaluation, Proceedings of international workshop on computational forensics (in IAS 2007) (2007) 411.

[27] C. E. H. Berger and D. Ramos, Ojective paper structure comparison: assessing comparison algorithms, Forensic Science International 222, (2012) 360.

[28] D. Ramos, Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems, Ph. D. Thesis Universidad autonoma de Madrid, Spain, 2007.

[29] D. Ramos and G. Zadora, Information-theoretical feature selection using data obtained by scanning electron microscopy coupled with and energy dispersive X-ray spectrometer for the classification of glass traces, Analytica Chimica Acta (2011).

[30] G. Zadora and D. Ramos, Evaluation of glass samples for forensic purposes - an application of likelihood ratios and an information-theoretical approach, Chemometrics and intelligent laboratory systems 102, (2010) 63.

[31] G. Zadora, A. Martyna, D. Ramos, and C. G. G. Aitken, Performance of likelihood ratio methods, in: Statistical analysis in forensic science: evidential value of multivariate physicochemical data, first ed., John Wiley and Sons, Chichester, UK, (2014), Chap. 6.

[32] R. Hirz and A. M. Rizzi, Simulation of the weathering of gasolines using gas chromatographic retention data, Journal of the Forensic Science Society 31, (1991) 309.

[33] A. Bolck, I. Alberink, Variation in likelihood ratios for forensic evidence evaluation of XTC tablets comparison, J. Chemom. 25 (2011) 41–49.

List of tables

Table 1. Compounds used in the 13 ratios, retention indices and values for F'

Ratio nr. in	Numerator		RI	Denom
Fig. 1	1	1,3,5-trimethylbenzene	954.0	1-methy
	2	2,2,4,5,6-pentamethylheptane	1026.8	1-methy propylb
	3	n-nonane	900	isoprop
	4	1,2,3-trimethylbenzene	1005.6	1-methy isoprop
	5	1-methyl-3-propylbenzene	1035.6	1,3-dim ethylbe
	6	1,2,4,5-tetramethylbenzene	1101.0	1,2,3,5- tetrame
	7	2-methylhexane	661.2	trans-1, dimethy
	8	1,3-diethylbenzene	1032.6	1,2-dim ethylbe
	9	2-methylheptane	760.4	trans 1,

			cyclohe
10	2,4-dimethylhexane	728.5	trans,ci
			trimeth
11	2,2,5-trimethylhexane	781.3	trans-1,
			ethylme
12	2,4-dimethylheptane	827.3	ethylbe
13	cis-1,4-dimethylcyclohexane	802.7	trimeth

Table 2. Overview of rates of misleading evidence for different-source comparisons (%) for the three LR methods.

LR method	*dist 13 ratios*	*trends evap*	*multivariate*
$1 < LR < 10$	0.337	0.110	0.188
$10 < LR < 100$	0.182	0.162	0.233
$100 < LR < 1000$	0.071	0.065	0.292
$1000 < LR < 10^4$	0.052	0.039	0.123
$10^4 < LR < 10^5$	0.006	0.006	0.045
total	0.649	0.383	0.882

Figures

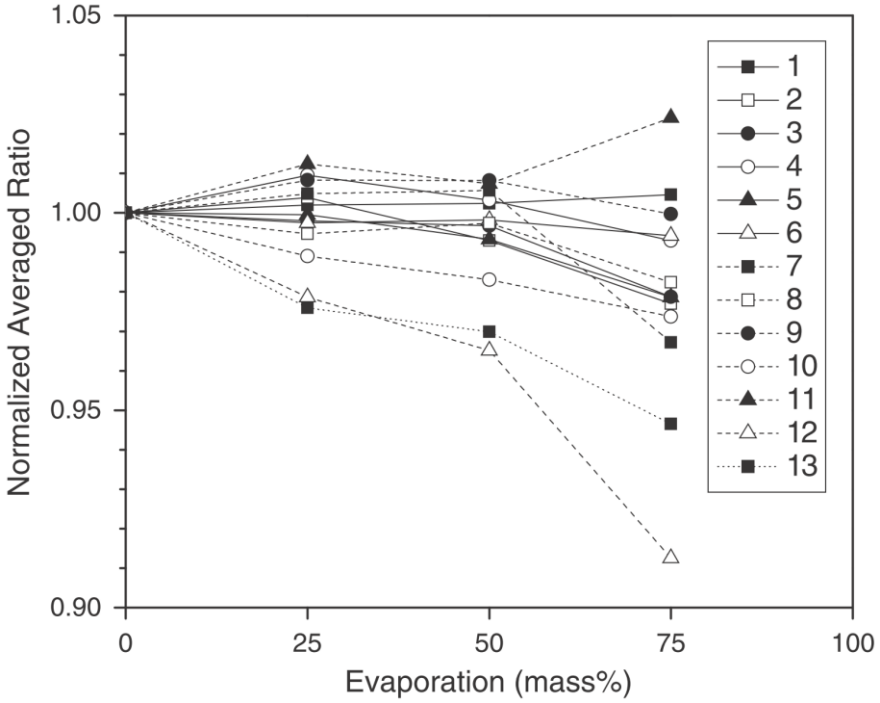

Fig. 1. Values of normalized (to 0% evaporation) ratios of peak areas in a chromatogram, averaged over gasolines and as a function of evaporation level.

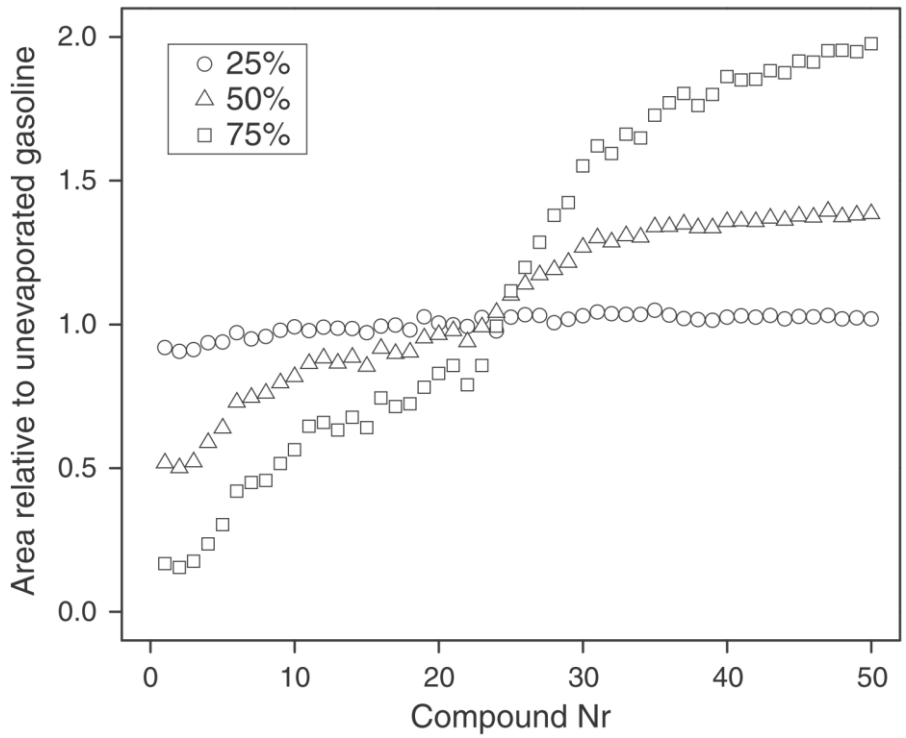

Fig. 2. Plot of relative peak areas in a chromatogram of 25, 50 or 75% evaporated gasoline, relative to peak areas in the chromatogram of the unevaporated gasoline. Compounds are ordered by increasing retention time.

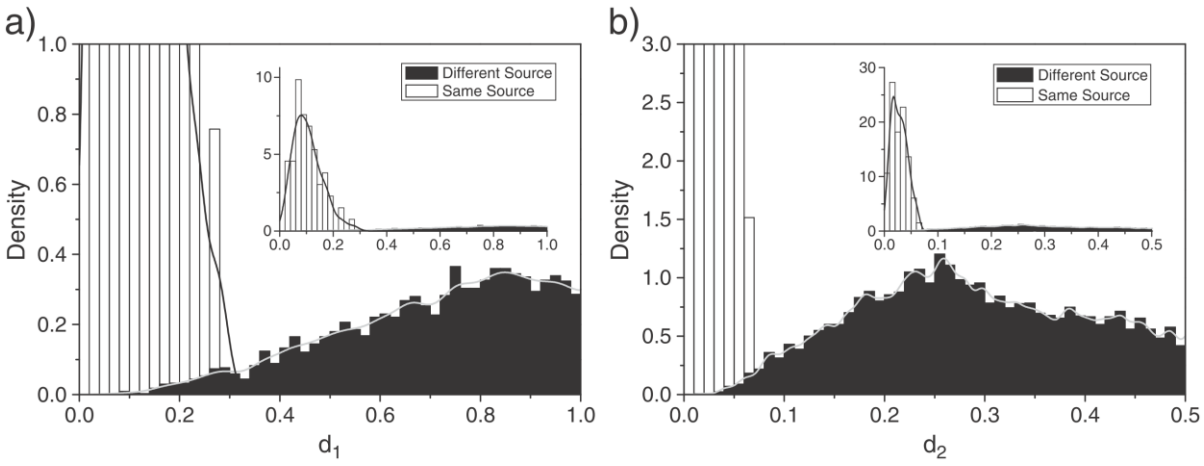

Fig. 3. Histograms of distances for the distance function based on (a) selected ratios and (b) trends in vapor pressure. The lines are the kernel smoothed densities.

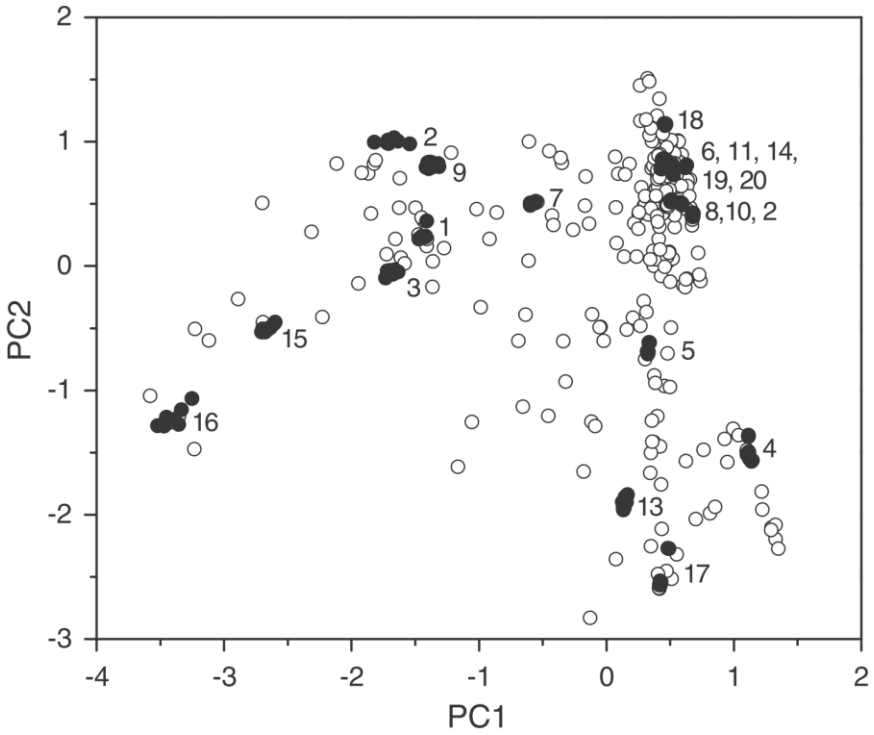

Fig. 4. PC scores plot for the validation dataset (black) and the background dataset (open circles) for the first two PCs.

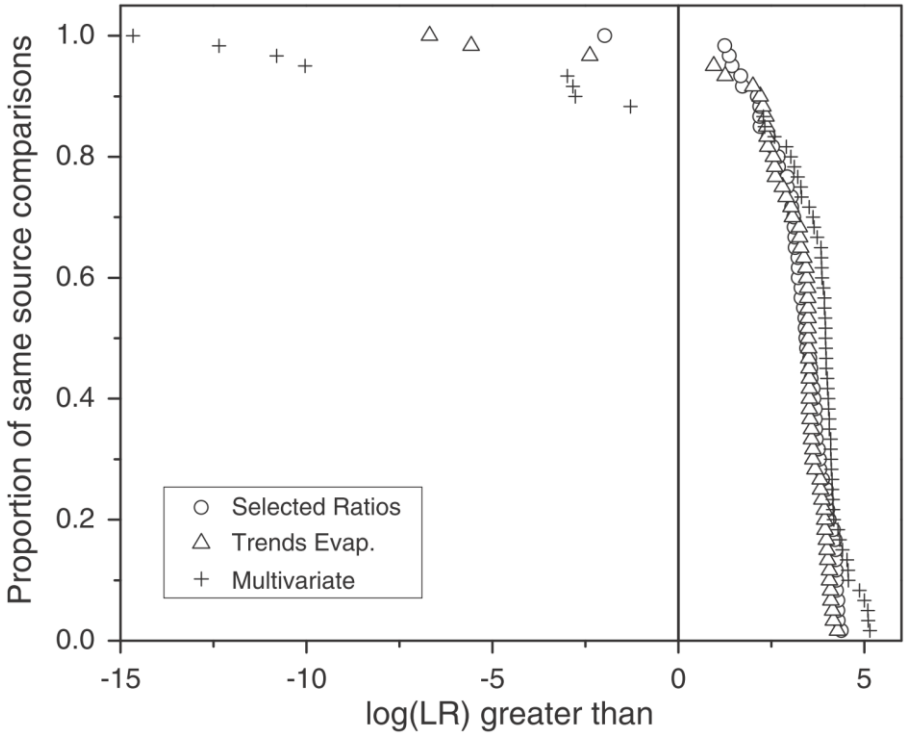

Fig. 5. A plot showing the proportion of same source comparisons with LRs larger than the value on the *x*-axis, for the three methods.

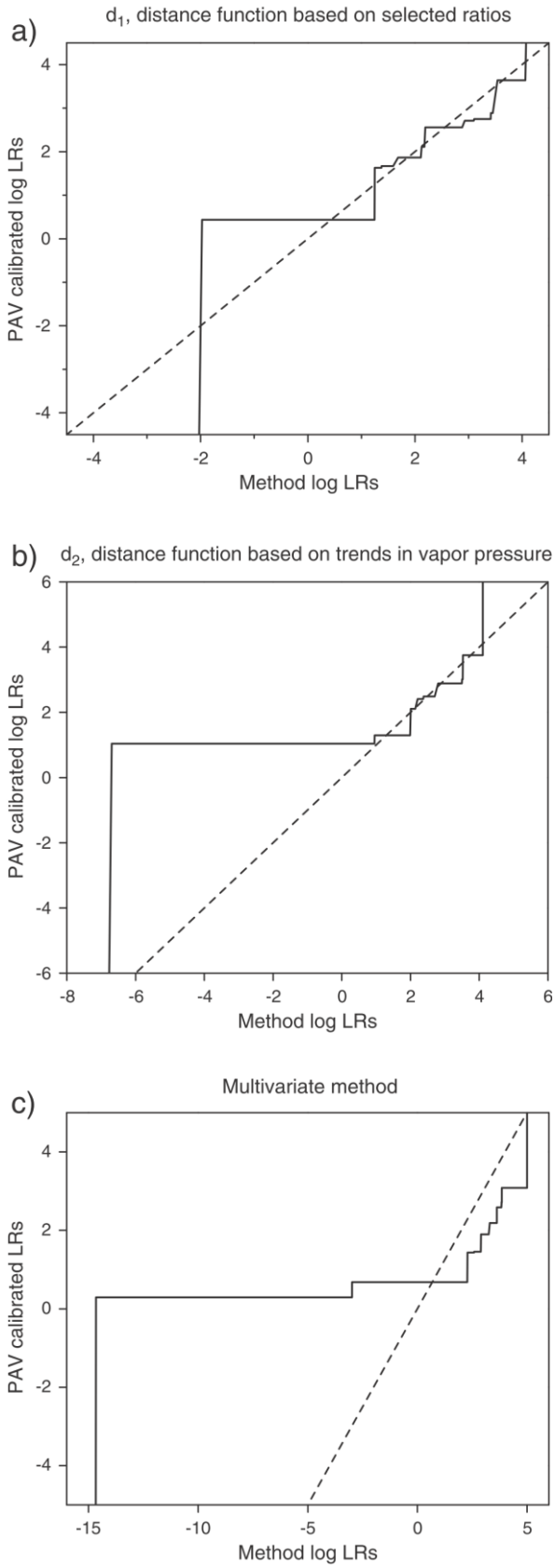

Fig. 6. PAV transforms of LRs of the validation data for (a) the distance function based on selected ratios, (b) the distance function based on trends in vapor pressure, and (c) the multivariate method.

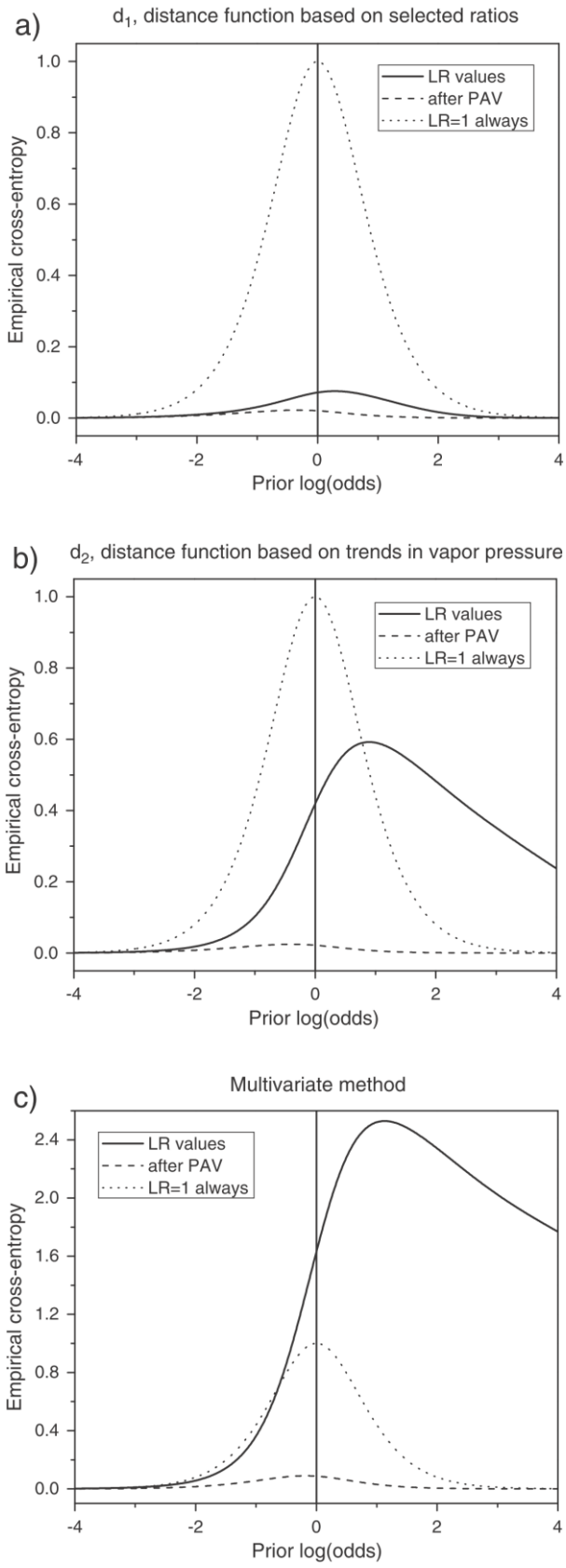

Fig. 7. ECE plots for (a) the distance function based on selected ratios, (b) the distance function based on trends in vapor pressure, and (c) the multivariate method.

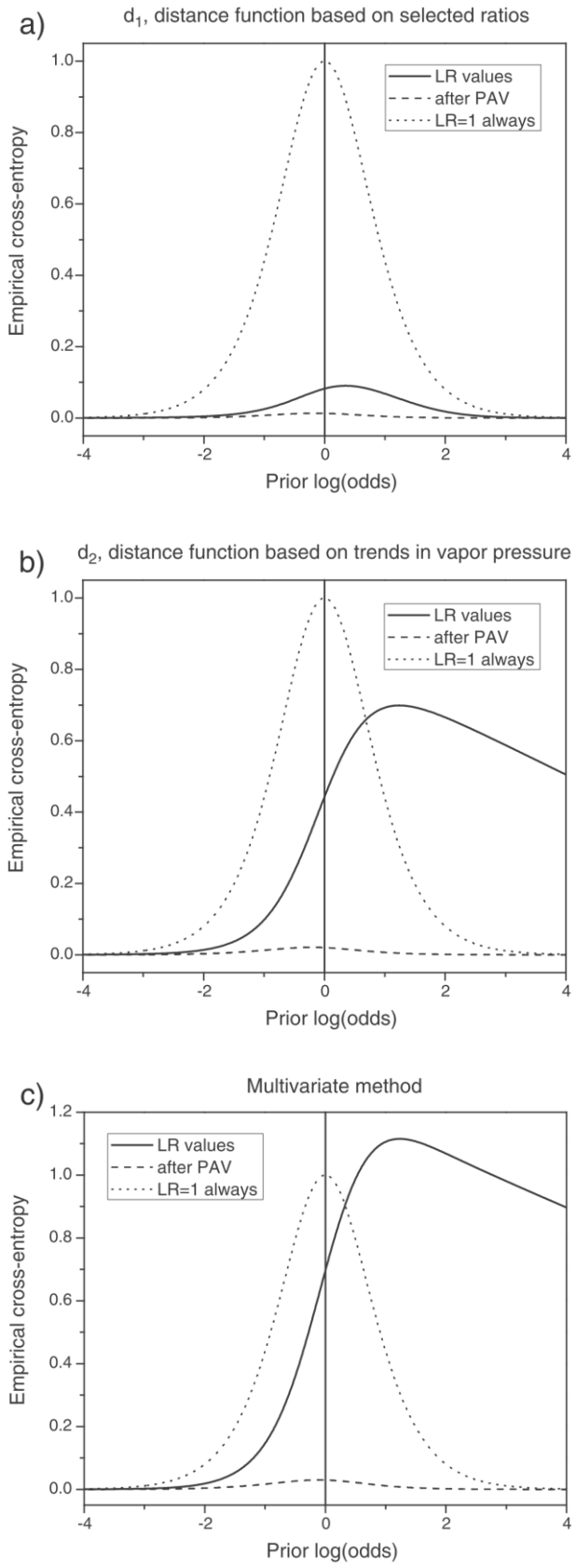

Fig. 8. ECE plots for LR-data of comparisons of gasoline residues that are both evaporated, (a) the distance function based on selected ratios, (b) the distance function based on trends in vapor pressure, and (c) the multivariate method.