



**Repositorio Institucional de la Universidad Autónoma de Madrid**

<https://repositorio.uam.es>

Esta es la **versión de autor** del artículo publicado en:

This is an **author produced version** of a paper published in:

Chemometrics and Intelligent Laboratory Systems 102.2 (2010): 63 – 83

**DOI:** <http://dx.doi.org/10.1016/j.chemolab.2010.03.007>

**Copyright:** © 2010 Elsevier

El acceso a la versión del editor puede requerir la suscripción del recurso

Access to the published version may require subscription

# Evaluation of glass samples for forensic purposes - an application of likelihood ratios and an information-theoretical approach

Grzegorz Zadora<sup>a</sup>, Daniel Ramos<sup>b</sup>

<sup>a</sup>*Institute of Forensic Research, Westerplatte 9, 31-033 Krakow, Poland.*

<sup>b</sup>*ATVS - Biometric Recognition Group, Universidad Autonoma de Madrid, 28049 Madrid, Spain.*

---

## Abstract

This article presents an experiential study about the influence of the selection of an adequate database for evidence evaluation using chemical profiles. Evidence evaluation in the forensic sense can be seen as the comparison of two glass objects, one of known origin, denoted as *control* glass, and typically would be from the scene of a crime, and the other one of unknown origin, termed *recovered* glass, which might be found in association with a suspect). The aim is to obtain some estimate of the weight of evidence for the degree of support to any of the hypothesis in the case, typically these might be that the *control and recovered glass come from the same source* ( $\theta_p$ ), and *control and recovered glass come from different sources* ( $\theta_d$ ). A likelihood ratio is considered a suitable measure of the evidential weight for the competing propositions. The observations are of the elemental composition of glass, measured using a Scanning Electron Microscopy, coupled with an Energy Dispersive X-ray spectrometer technique. A number of glass objects have been analyzed and their chemical profiles form a database which represents several sources of variation. In this paper questions surrounding the choice of observations to make are addressed empirically by assessing the impact of building each model using a database different from the one using for comparison. The performance of each evidence evaluation method is assessed by classical methods such as Tippett plots, or more recent information-theoretical approaches such as empirical cross-entropy (ECE) plots.

The results show that several of the compositional elements are very robust to the selection of the background database, namely; calcium, silicon

and sodium observed in their oxide forms. We also show that the likelihood ratio computed with the combination of these variables show a remarkable discriminating power, and good calibration, allowing them to be employed for the calculation of the strength of evidence in forensic case work.

*Key words:*

Forensic science, evidence evaluation, likelihood ratio models, empirical cross entropy, physico-chemical multivariate data, glass.

---

## 1. Introduction

The importance of glass as evidence in forensic cases has noted for many years. One of the properties which makes it suitable for forensic case work is the generation of very small glass fragments (0.1 - 0.5 mm) that arise during car accidents, burglaries, fights *etc.* which can be transferred to, and carried on, the clothes, shoes, and hair of participants [1]. Because of their very small size glass fragments are analysed by analytical methods which give reliable observations for small objects. The GRIM (Glass Refractive Index Measurement) method (*e.g.* in [1]) and Scanning Electron Microscopy coupled with an Energy Dispersive X-ray spectrometer (SEM-EDX), are routinely used in many forensic institutes for the investigation of glass and other forensic problems [2]. Other methods of elemental analysis of glass fragments are  $\mu$ -X-Ray Fluorescence [3] and Laser Ablation-Inductively Coupled Plasma-Mass Spectrometry [4]. However, these methods require relatively large fragments of glass. For example LA-ICP-MS gives good results with pieces of glass larger than 0.5 mm. SEM-EDX has the drawback that it can only provide information about the major and minor elements such as; O, Na, Al, Mg, Si, K, Ca, Fe. It has been believed that trace element concentrations are essential to enable the glass investigator to effectively compare and individualise glass evidence, however, it has been shown that some headway can be made on the basis of the major and minor element concentrations [5, 6, 7, 8].

Standard modern forensic practice is to interpret physicochemical data as evidence ( $E$ ), in the context of two competing hypotheses or propositions, one proposed by prosecutor, and one proposed by defence. Consider a case where the fact finder, that is a prosecutor, or a judge, asks a forensic scientist to evaluate the evidential value of a recovered glass fragment of unknown

origin, and a control glass fragment, whose origin is known. The relevant propositions for the fact finder arise from the circumstances of the case, and often because of the adversarial nature of the system, but for evidence evaluation they typically are:

- $\theta_p$ : the control and recovered glass fragment come from the same glass object. This is often called the *prosecutor's proposition*,
- $\theta_d$ : the control and recovered glass fragment come from different glass objects from the same overall population of glass objects. This can be called the *defence proposition*.

The task set for the forensic scientist can be termed a comparison problem, and it requires careful attention to the following considerations:

1. the similarity of physicochemical features determined for the recovered material to the control sample.
2. some estimate of the rarity of the determined physicochemical features for recovered and control glass samples in the relevant population;
3. possible sources of uncertainty (sources of error) which will include, at least: variation of measurements of physicochemical features within the recovered and control glass fragments: and variation of measurements of physicochemical features between various glass objects in the relevant population,
4. the possible correlations between different physicochemical features when more than one physicochemical feature has been measured.

Attempts to solve comparison problem have been based upon a classical approach, or a Bayesian approach. The classical approach uses significance tests, a set of ideas developed in the first half of the twentieth century. The approach involves two-stages. First, there is a comparison stage, and then a rarity stage. In the first stage, evidence from the crime is compared with evidence from the environment of the suspect. There is a binary outcome to this comparison. The two sets of evidence may be deemed similar or dissimilar. If they are deemed similar, the second stage of the analysis is conducted. If they are deemed dissimilar then the analysis is stopped and it is decided to act as if the two sets of evidence came from different sources. For those sets of evidence deemed similar, the second stage is the assessment of the rarity of the evidence. Similarity of two sets of evidence which are rare

in the characteristics of similarity (*e.g.* physicochemical data) will be deemed stronger evidence of a common source for the two sets than a similarity of two sets of evidence which are common in the characteristics of similarity.

More recent approaches based in the Bayesian paradigm have been applied to comparison problems. Various approaches can be employed, however, those related to the likelihood ratio are considered the best documented, and most developed, measure of evidential value for forensic purposes [10]. A key concept in such approaches is the law of likelihood, which is stated as follows: if  $\theta_p$  implies that the probability that a random variable  $X$  takes the value  $x$  is  $Pr(X = x|\theta_p)$ , while hypothesis  $\theta_d$  implies that the probability is  $Pr(X = x|\theta_d)$ , then the observation  $X = x$  is evidence supporting  $\theta_p$  over  $\theta_d$  if and only if  $Pr(X = x|\theta_p) > Pr(X = x|\theta_d)$ , and the likelihood ratio  $LR = Pr(X = x|\theta_p)/Pr(X = x|\theta_d)$ , measures the strength of that evidence [11].

Bayes' theorem [10] relates conditional probabilities and their transposes:

$$Pr(\theta_p | E, I) = \frac{Pr(E | \theta_p, I) \cdot Pr(\theta_p | I)}{Pr(E | I)} \quad (1)$$

where  $I$  is the background information available in the case, and is not related to the evidence  $E$ . This  $I$  may include not only circumstantial information in the case (such as witness testimony), but also the analysis of other forensic evidence apart from  $E$  (such as other glass fragments, paint flakes, etc.). Equation 1 then allows the following inference:

$$\frac{Pr(\theta_p | E, I)}{Pr(\theta_d | E, I)} = LR \cdot \frac{Pr(\theta_p | I)}{Pr(\theta_d | I)} \quad (2)$$

$$LR = \frac{Pr(E | \theta_p, I)}{Pr(E | \theta_d, I)} \quad (3)$$

Equation 2 is the *odds* form of Bayes' theorem, where the term *odds* refers to as the quotient of two complementary probabilities. The hypotheses can be defined for the court from  $I$ , the prosecution and defense propositions. In this framework, we can distinguish two values:

1. The prior probabilities  $Pr(\theta_p | I) = 1 - Pr(\theta_d | I)$ , which are province of the fact finder, and should be stated assuming only the background information ( $I$ ) in the case [12].

2. The LR (Equation 3<sup>1</sup>), computed by the forensic scientist [10].

Values of LR above 1 support  $\theta_p$  and values of LR below 1 support  $\theta_d$ . A value of LR close to 1 provides little support for either proposition. Also the larger (the lower) the value of the LR, the stronger (the weaker) the support of the evidence for  $\theta_p$  ( $\theta_d$ ). The most popular application of the LR approach for forensic purposes is the evaluation of the evidential value of matching DNA profiles. Likelihood ratios have been also used in earprints, fingerprints, firearms and toolmarks, speaker recognition, hair, documents and handwriting [10]. There are also an increasing number of applications of this approach in evaluation of physicochemical data for univariate or multivariate data [8, 13], in particular where the observations of elemental composition were by means of SEM-EDX [5, 6, 7, 8] or paint data [14].

More generally likelihood ratio based approaches could be used in many situations where propositions are compared, and are especially relevant where a decision made on the basis of physicochemical analysis may have serious legal or economical consequences. For instance, a likelihood ratio could be used when food is analysed and the questions revolve around whether the food item in question contains what is described on the label ( $\theta_p$ ) or not ( $\theta_d$ ). In these sorts of instance it might be more appropriate to use a likelihood ratio based approach than a classical frequentist approach, particularly when any classical test employed might suggest a value close to the threshold value for significance. In such situations classical approaches are burdened by a fall-off-cliff problem. For instance, for  $\alpha = 0.05$ , the difference between a comparison which is significant at the 0.051 level, and a difference which is significant at the 0.049 level, has a very large effect on the decision making. This problem is reduced by a likelihood ratio approach, because if the data after analysis is close to the threshold (“the edge of a cliff”) then the LR value should be close to 1. In such a situation it suggests that any decision as to which hypothesis is correct should be made with regard to other factors, as the observations provide little support one way or the other.

One fundamental problem in evidence evaluation using likelihood ratios is related to population databases. The evaluation of evidential value requires an assessment of the rarity of the evidence. This stage requires information collected in database about various similar objects, and known as a *popula-*

---

<sup>1</sup>The background information  $I$  is always conditioning all the probabilities, and will be eliminated from the notation for simplicity.

tion. This information is also used to calculate the between-object variability. The selection of a proper database (a *relevant* population) is one of the crucial factors in the evaluation of evidential value of physicochemical data. This relevant population database is also known as background data, and it is essential for tuning the models used in LR computation. Glass objects to be analysed in daily casework may have several different origins. In this work glass objects coming from car windows (*cw*), building windows (*bw*) and containers (*p*) are considered. Glass objects from *cw* and *bw* categories have very similar elemental compositions, especially information about the main elemental content determined by SEM-EDX, because they are manufactured in very similar way (a float glass manufacturing method), and they were grouped in the presented research into the same category (windows, *w*). However one might expect that container (*p*) glasses have a systematically different elemental composition than windows (*w*), due to the different type of manufacturing process and due to their different use purpose. If a LR value is computed for control and recovered samples of container *p* glass using the models based upon a background sample comprising window glass, it is intuitively reasonable for those LR values to be less accurate than for likelihood ratio calculations based upon observations including window glass.

This paper focuses on describing to what extent such variation may impact the performance of the computed LR values. Another problem examined in this paper is influence on likelihood ratios for comparison problems of the observed variability of the elemental composition calculated for samples of building window glass from Polish and British contexts. An additional contribution in this work is the comparison of several assessment metrics for LR-based evidence evaluation. The aim is to evaluate the impact in the performance of the process among different experimental scenarios by the use of three assessment methods: first, we evaluate the *rates of misleading evidence* of the experimental LR values obtained, which have been dubbed *false positive* and *false negative* rates in previous literature [10]. Second, we use Tippett plots [15], which are popular in several disciplines in forensic interpretation of the evidence in the form of LR values. Finally, we present Empirical Cross-Entropy (ECE) plots as a measure of the information given by the evidence evaluation process. As it will be explained later, the latter is based on previous work in the statistical literature based on strictly proper scoring rules [16, 17], and has been proposed as a proper way of assessing the performance of LR values in different disciplines in forensic science [18, 19, 20].

## 2. Methods and models

### 2.1. Glass databases and experimental protocol

#### 2.1.1. Polish database

The database used in this paper consists of 222 glass-objects, which include 57 containers ( $p$ ), 165 float glass ( $w$ , *i.e.* 79 building windows ( $bw$ ) and 86 car windows ( $cw$ )).

#### 2.1.2. British database

Comprise 82 float glass objects ( $w$ ) originating from building windows.

### 2.2. Elemental composition analysis by SEM-EDX technique

Four glass fragments, with surfaces as smooth and flat as possible, collected from each glass object, were placed on self-adhesive carbon tabs on an aluminium stub, and then carbon coated using an SCD sputter (Bal-Tech, Switzerland). Three replicate measurements on each fragment were made. Analysis of the elemental content of each glass fragment was carried out using a Scanning Electron Microscope (JSM-5800 Jeol, Japan), with an Energy Dispersive X-Ray detector (Link ISIS 300, Oxford Instruments Ltd., UK). The measurement conditions were: accelerating voltages  $20kV$ , life time  $50s$ , magnification  $1000 - 2000\times$ , and the calibration element was cobalt. The SEMQuant option (part of the software LINK ISIS, Oxford Instruments Ltd, UK) was used in the process of measuring the percentage of particular elements in a fragment. The selected analytical conditions allowed the estimation of concentrations of oxygen, sodium, magnesium, aluminium, silicon, potassium, calcium and iron. The data consist of seven variables obtained as the  $\log_{10}$  of the ratio with respect to the oxygen (O) concentration, which is always non-zero for glass data. The seven variables thus obtained will be denoted as ( $Na'$ ,  $Mg'$ ,  $Al'$ ,  $Si'$ ,  $K'$ ,  $Ca'$ ,  $Fe'$ ). When 0 was observed in the raw data it was substituted by a small value (0.0001).

### 2.3. Likelihood ratio model

A likelihood ratio model which considers two levels of uncertainty, within and between object variability, for multivariate data has been used in addressing comparison problems has been proposed in [6] and successfully tested on data obtained for glass objects analysed with SEM-EDX [5, 6, 7, 8] or paint data [14].



A population with  $p$  characteristics is assumed and that a background database is available with  $n$  observations on each of  $m$  objects, so that data are in the form of  $p$ -vectors  $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})^T$  for  $(i = 1, \dots, m)$  and  $(j = 1, \dots, n)$ . The object means of these measurements are  $\bar{\mathbf{x}}_i = \sum_{j=1}^n \mathbf{x}_{ij}/n$ .

Suppose there are two sets, one of  $n_c$ , one of  $n_r$ , measurements made respectively for control and recovered items on the  $p$  characteristics, *i.e.* respectively  $\mathbf{y}_{cj} = (y_{cj1}, \dots, y_{cjp})^T$ ,  $\mathbf{y}_{rj} = (y_{rj1}, \dots, y_{rjp})^T$ . The means of the control and recovered samples are respectively  $\bar{\mathbf{y}}_c = \sum_{j=1}^{n_c} \mathbf{y}_{cj}/n_c$  and  $\bar{\mathbf{y}}_r = \sum_{j=1}^{n_r} \mathbf{y}_{rj}/n_r$ .

It is also assumed that the within-object distribution is multivariate normal with constant variance from object to object, since it is difficult to estimate separate variances for each object due to the small number of observations per object. The within-object variance matrix  $\mathbf{U}$  is estimated as:

$$\mathbf{U} = \frac{\mathbf{S}_w}{mn - m} \quad \text{where} \quad \mathbf{S}_w = \sum_{i=1}^m \sum_{j=1}^n (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T.$$

The between-object variance matrix  $\mathbf{C}$  is estimated as:

$$\mathbf{C} = \frac{\mathbf{S}^*}{m - 1} - \frac{\mathbf{S}_w}{n(nm - m)} \quad \text{where} \quad \mathbf{S}^* = \sum_{i=1}^m (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T.$$

Early work assumed that the between-object distribution is multivariate normal, however, this assumption is unlikely to be true in any real population of glass objects. Therefore, a LR model used in this paper a kernel density approach was adopted for modelling the between-object distributions using Gaussian kernels. Then the between-source distribution is estimated from the group means using a multivariate kernel density function with normal kernel functions having each one mean  $\bar{\mathbf{x}}_i$  and covariance  $h^2\mathbf{C}$  [6].

The numerator of the likelihood ratio, for which  $\theta_p$  is assumed true, can be shown (e.g. see Appendix in [9]) to be given by:

$$\begin{aligned} (2\pi)^{-p/2} \left| \frac{\mathbf{U}}{n_c} + \frac{\mathbf{U}}{n_r} \right|^{-1/2} & \exp\left\{-\frac{1}{2}(\bar{\mathbf{y}}_c - \bar{\mathbf{y}}_r)^T \left(\frac{\mathbf{U}}{n_c} + \frac{\mathbf{U}}{n_r}\right)^{-1} (\bar{\mathbf{y}}_c - \bar{\mathbf{y}}_r)\right\} \\ & \times \frac{1}{m} \sum_{i=1}^m (2\pi)^{-p/2} \left| \frac{\mathbf{U}}{n_c + n_r} + h^2\mathbf{C} \right|^{-1/2} \\ & \times \exp\left[-\frac{1}{2}(\bar{\mathbf{y}} - \bar{\mathbf{x}}_i)^T \left(\frac{\mathbf{U}}{n_c + n_r} + h^2\mathbf{C}\right)^{-1} (\bar{\mathbf{y}} - \bar{\mathbf{x}}_i)\right] \end{aligned}$$

where  $\bar{\mathbf{y}}$  is the overall mean of the control and recovered measurements:

$$\bar{\mathbf{y}} = \frac{n_c \bar{\mathbf{y}}_c + n_r \bar{\mathbf{y}}_r}{n_c + n_r}.$$

The denominator, for which  $\theta_d$  is assumed true, can be shown (e.g. see Appendix in [9]) to be given by:

$$\begin{aligned} & (2\pi)^{-p/2} \left| \frac{\mathbf{U}}{n_c} + h^2 \mathbf{C} \right|^{-1/2} \times \frac{1}{m} \sum_{i=1}^m \exp -\frac{1}{2} (\bar{\mathbf{y}}_c - \bar{\mathbf{x}}_i)^T \left( \frac{\mathbf{U}}{n_c} + h^2 \mathbf{C} \right)^{-1} (\bar{\mathbf{y}}_c - \bar{\mathbf{x}}_i) \\ & \times (2\pi)^{-p/2} \left| \frac{\mathbf{U}}{n_r} + h^2 \mathbf{C} \right|^{-1/2} \times \frac{1}{m} \sum_{i=1}^m \exp -\frac{1}{2} (\bar{\mathbf{y}}_r - \bar{\mathbf{x}}_i)^T \left( \frac{\mathbf{U}}{n_r} + h^2 \mathbf{C} \right)^{-1} (\bar{\mathbf{y}}_r - \bar{\mathbf{x}}_i) \end{aligned}$$

According to [21], an optimal value,  $h_{\text{opt}}$ , for the window smoothing parameter  $h$  for the kernel distribution is expressed as:

$$h = h_{\text{opt}} = \left( \frac{4}{2p+1} \right)^{\frac{1}{p+4}} \frac{1}{m^{\frac{1}{p+4}}}.$$

### 2.3.1. Graphical model

A limitation of multivariate modelling is the lack of background data from which to estimate the parameters of the assumed distributions such as means, variances and covariances. For example, if objects are described by  $p$  variables then it is necessary to reliably estimate up to  $p$  means,  $p$  variances, and  $p \cdot (p-1)/2$  covariances in LR model presented above if someone would like to use all  $p$  variables in LR calculations in the form:

$$LR = \frac{(Var_1, \dots, Var_p \mid \theta_p)}{(Var_1, \dots, Var_p \mid \theta_d)}$$

Thus this process requires far more data than are usually available in many forensic databases. This effect has been dubbed the ‘curse of dimensionality’.

The simplest way around this problem is to assume independence of the variables considered, however this is in most of cases a naive assumption as by nature the measured features are correlated, especially those pertaining to the main components of the analysed materials. Where feasible, models should take into account this correlation, possibly in combination with some

form of dimension reduction, *e.g.* graphical models [22], and the independence assumption should be used only when the number of observations is extremely limited (and taken care of its implications). Graphical models are a dimension reduction technique based on a graph theory.

A graphical model is selected by the sequential addition of edges decided by inspection of the negative partial correlation coefficient matrix (scaled inverse correlation matrix  $\mathbf{C}^{-1}$ ). For illustration, let us consider the 3-variable problem for which the following values of partial correlation coefficients were obtained:  $-0.1$  between A and B,  $0.8$  between B and C and  $0.6$  between A and C. First, the largest magnitude partial correlation is selected, and an edge is added between the two nodes connected by this partial correlation. In this case these are respectively  $-0.8$ — and B-C. This process is repeated until all nodes are part of the graphical model (Fig. 1) or the model can not be factorised [22]. The factorisation of the  $p$ -dimensional density (here: 3-dimensional) one is given by:

$$f(C_i|S_i) = \frac{f(C_i)}{f(S_i)}$$

where  $C_i$  is the  $i$ th clique in the model, *i.e.* a subset of variables in which all the nodes are connected to each other is known as a complete subgraph, and  $S_i$  is the set of all separators for the  $i$ -th clique, *i.e.* a node or a set of nodes in the intersection of two cliques. In the case of a simple model presented in Fig. 1 two are two cliques (A, C) and (B, C) and one separator, *i.e.* a node C. Finally, the following factorisation is obtained:

$$f(C_i|S_i) = \frac{f(A, C)f(B, C)}{f(C)}$$

This factorisation of the density corresponds to the following factorisation for the LR models:

$$LR = \frac{LR(A, C)LR(B, C)}{LR(C)}$$

where  $LR(A, C)$  is the value of the likelihood ratio calculated based on multivariate densities with three variables (A, C) by application of the LR model described in point 2.3 and so on.

In the case of more sophisticated graphical models a procedure to find cliques and separators is described in [6]. Other examples of application of

the graphical models approach for dimension reduction evaluation of evidence value of physicochemical data can be found in [5, 8, 7, 23].

Various experiments were performed in the presented research and each one considered different glass databases. Thus, different  $\mathbf{C}$  matrices were obtained in each of these experiments. That yielded some differences in the constructed graphical models. It was not possible to conclude which of obtained graphical models should be used as they were built on the base of results of analysis of relatively limited number of glass data, *i.e.* maximum 165 float glass samples. Therefore, the following *standardised* graphical model was used, with the aim of eliminating the influence of the application of various graphical models on the base of the obtained results. It was constructed on authors experiences from previous research and knowledge chemical composition of glass, and this is a simplified version of a previously obtained graphical model for glass data [6]. The proposed model accounts for the main relationships between the most important chemical compounds in glass. The 7 variables were divided into 4 cliques (Na', Si', Ca'), (Al', K'), (Mg') and (Fe'). The clique (Na', Si', Ca') involves elements which form the basis of soda-lime-silica glass. A high correlation between Al' and K' was also observed in previous analyses. These variables represent compounds used in glass as additives whose role is to improve the physical properties of glass samples (optical -  $K_2O$  - or to avoid re-crystallisation process -  $Al_2O_3$ ). The variables Mg' and Fe' could be treated as independent variables. Finally, the LR was calculated by the following formula:

$$LR = LR(Na', Si', Ca')LR(Al', K')LR(Mg')LR(Fe') \quad (4)$$

where  $LR(Na', Si', Ca')$  is the value of the likelihood ratio calculated based on multivariate densities with three variables (Na', Si' and Ca') by application of the LR model described in point 2.3;  $LR(Al', K')$  is the value of the likelihood ratio calculated based on multivariate densities with two variables (Al' and K'); and similarly for  $LR(Mg')$ ,  $LR(Fe')$  and so on.

The likelihood ratios obtained for pairwise comparisons from the *standardised* graphical model were compared with similar comparisons from the graphical model published in [6] and, as well as giving very good results against known outcomes, has an easily interpretation from a chemical point of view. This model used allowed a the 7-dimensional problem to be condensed into sets of 2-dimensional, and univariate problems and the final formula for the calculation of the LR was:

$$LR = \frac{LR(Na', Si')LR(K', Na')LR(Al', Fe')LR(Al', K')LR(Ca', Na')LR(Ca', Mg')}{LR(Na')^2LR(K')LR(Al')LR(Ca')} \quad (5)$$

This model is used in routine casework by one of the authors.

#### 2.4. Experimental protocol

For each set of measurements of glass considered as *control* data, and each set of measurements considered as *recovered* data, a LR will be computed. If the set of control and recovered data comes from the same source ( $\theta_p$  is true) we will refer the resulting LR to as a *same-source* comparison or a *same-source* LR value. On the other hand, if the set of control and recovered data comes from different sources ( $\theta_d$  is true) we will refer the resulting LR to as a *different-source* comparison or a *different-source* LR value. Thus, performing all the possible pair-wise comparisons among each set of control and recovered data, an experimental set of LR values will be obtained, with  $N_p$  same-source LR values and  $N_d$  different-source LR values. This experimental set will be used for performance assessment.

#### 2.5. Performance assessment and representation

In this work, we use three methods for the assessment of the performance of LR values: the rates of misleading evidence, Tippett plots and ECE plots. These assessment methods are based on the experimental set of LR values generated with the database and experimental protocols described in 2.4.

##### 2.5.1. Rates of misleading evidence

The LR value supports either the prosecution, or defense hypotheses  $\theta_p$  or  $\theta_d$ , as described in Sections 1 and 2.3. However, the support given by a single LR value can be misleading due to the imperfect nature of the evidence evaluation process. The ways in which this may happen are the following:

- The LR value is lower than 1 and the true value of the hypothesis is  $\theta_p$ . The rate of LR values for which this error occurs defines the rate of misleading evidence when  $\theta_p$  is true. If this error occurs, the LR value is giving support to the  $\theta_d$  hypothesis when  $\theta_p$  is true. This kind of error has been referred to as *false negative rate* [10].

- The LR value is greater than 1 and the true value of the hypothesis is  $\theta_d$ . The rate of LR values for which this error occurs defines the rate of misleading evidence when  $\theta_d$  is true. If this error occurs, the LR value is giving support to the  $\theta_p$  hypothesis when  $\theta_d$  is true. This kind of error has been referred to as *false positive rate* [10].

Although this metric is illustrative, it gives only a little information about the performance of the LR values, because the concept of misleading evidence implicitly considers only a threshold of  $LR = 1$  for determining the goodness of the experimental set. However, they are insensitive to, *e.g.*, whether a comparison where  $\theta_d$  is true yields a LR value of 10 or 1000, which would indicate much strongly misleading support to the wrong hypothesis  $\theta_p$  in the latter. Such effect is ignored by the rates of misleading evidence, but it is critical for the performance of LR values.

### 2.5.2. Tippett plots

Tippett plots have been classically used for empirical performance assessment [15], and consist of the cumulative distributions of LR values for same-source, and different-source experiments. In particular, two curves are plotted, one for same-source LR values ( $\theta_p$  is true) and one for different-source LR values ( $\theta_d$  is true). For each curve, and for each value  $x_0$  in the logarithmic  $x$  axis, the proportion of LR values which are greater than  $x_0$  is represented. It is worth noting that Tippett plots include the values of the rates of misleading evidence, which are determined by the value of each curve at  $x_0 = 0$ . An example of Tippett plots is shown in Figure 2, where the rates of misleading evidence are highlighted. This figure shows also an example of how Tippett plots yield more information about performance than the rates of misleading evidence. The figure shows the Tippett plots of two sets of LR values, namely  $A$  and  $B$ . It is shown that LR set  $A$  shows some given values of misleading evidence and a limited value of misleading LR values, which is identified as a desirable behavior. However, LR set  $B$  shows a significant proportion of strongly misleading LR values, even though the rates of misleading evidence in set  $A$  and set  $B$  are extremely similar. Therefore, in this case rates of misleading evidence are useless to detect the degrading effects of strong misleading evidence, which are effectively seen in Tippett plots.

Although Tippett plots are useful and clearly show many important performance indicators for the given experimental set of LR values, we identify some problems for this kind of performance representation. In particular,

Tippett plots do not give a scalar number indicating which of the two sets of LR values is better, and in consequence sometimes it is difficult to state which evidence evaluation method outperforms which other one. Another problem is that the impact of misleading evidence is not explicitly measured; although it is known that LR values  $\ll 1$  when  $\theta_p$  is true or  $\gg 1$  when  $\theta_d$  is true are undesirable, in Tippett plots their impact is not numerically measured. In addition to these problems, Tippett plots do not explicitly measure the discriminating power, understood as the degree of *separation* among same-source and different-source LR values. In Tippett plots, the discriminating power can be viewed as the comparison between the false positive and false negative rates at any given value  $x_0$  of  $\log_{10}(LR)$ , but an overall measure of discriminating power is not explicitly given.

### 2.5.3. Empirical Cross-Entropy (ECE) plots

In order to solve the drawbacks in the use of Tippett plots and rates of misleading evidence, we propose an assessment metric based on information theory [24], namely Empirical Cross-Entropy (ECE) [18, 20]. The motivation of such a measure is derived from the statistics literature. In [16], the performance of an experimental set of posterior probabilities, expressed as a forecast from a given forecaster, is assessed by means of *strictly proper scoring rules*. An example of strictly proper scoring rule, which we will use in this work, is the logarithmic scoring rule. For each value of the evidence  $E$  in a forensic case, the logarithmic scoring rule takes the following values<sup>2</sup>:

$$\begin{aligned} \theta_p \text{ true} &: -\log_2(Pr(\theta_p|E)) \\ \theta_d \text{ true} &: -\log_2(Pr(\theta_d|E)) \end{aligned} \tag{6}$$

Thus, strictly proper scoring rules may be viewed as loss functions which assign a penalty to a given value of the posterior probability depending on: *i*) the value of the posterior probability, and *ii*) the true hypothesis among  $\theta_p$  and  $\theta_d$  which actually occurred [16, 17]<sup>3</sup>. For example, if a probabilistic forecast gives a high posterior probability of raining tomorrow (value of the forecast) and tomorrow it does not rain (true hypothesis), a strictly proper scoring rule will assign a high penalty to the forecast, and vice-versa. The logarithmic scoring rule is illustrated in Figure 3. In [16] the overall measure

---

<sup>2</sup>We take base-2 logarithms for normalization of the derived ECE value, but the base is irrelevant for comparison of performance, since it just represents a global scaling factor.

<sup>3</sup>See [16, 17] for more examples of strictly proper scoring rules

of goodness of a forecaster is defined as the average value of a strictly proper scoring rule over many different forecasts. For instance, for the logarithmic scoring rule, this mean value would be the *logarithmic score* ( $LS$ ):

$$LS = - \frac{1}{N_p} \sum_{i \in \text{targets}} \log_2 Pr(\theta_p | e_j) - \frac{1}{N_d} \sum_{j \in \text{nonTargets}} \log_2 Pr(\theta_d | e_j) \quad (7)$$

Where  $N_p$  and  $N_d$  is the number of same-source and different-source LR values being evaluated. This average value  $LS$  can be viewed as an overall *loss*. Moreover, it is also demonstrated in [16] that such a measure of performance can be divided into two components:

1. A *calibration loss* component, which measures how similar are the forecasts to the frequency of occurrence of  $\theta_p$ . Low calibration loss means that for a given range of values of the forecast  $Pr(\theta_p | E)$  closely around  $x$ , the frequency of cases where actually  $\theta = \theta_p$  tends to be  $x$ .
2. A *refinement loss* component, which measures how sharp or how spread the forecasts are. Roughly speaking, low refinement loss means that if the calibration loss of the forecaster is low, for a given value of the forecast  $Pr(\theta_p | E)$  the frequency of trials where actually  $\theta = \theta_p$  is near 0 or 1.

The proposed measure of goodness is a variant of  $LS$ , and is expressed as follows:

$$ECE = - \frac{Pr(\theta_p)}{N_p} \sum_{i \in \text{same-source}} \log_2 Pr(\theta_p | e_j) - \frac{Pr(\theta_d)}{N_d} \sum_{j \in \text{diff-source}} \log_2 Pr(\theta_d | e_j) \quad (8)$$

where same-source and different-source refers to the comparisons in the experimental set of LR values when  $\theta_p$  and  $\theta_d$  are respectively true, and  $N_p$  and  $N_d$  are the number of such respective comparisons as explained in Section 2.4.



ECE has an attractive and simple interpretation: the higher its value, the more the information the fact finder needs in order to know the true value of the hypotheses. This information should be interpreted on average over different forensic cases (comparisons in the experimental set), and not the information given by a particular case, where the true hypothesis is generally not known. If the LR values of the evidence evaluation process are misleading to the fact finder, then the ECE will grow, and more information on average will be needed in order to know the true values of the hypotheses. The details about the derivation and interpretation of ECE can be found in [18].

As it can be seen, ECE depends on the prior probability  $Pr(\theta_p) = 1 - Pr(\theta_d)$ , since:

$$ECE = \frac{Pr(\theta_p)}{N_p} \sum_{i \in \text{same-source}} \log_2 \left( 1 + \frac{1}{LR_i \cdot \frac{Pr(\theta_p)}{Pr(\theta_d)}} \right) + \frac{Pr(\theta_d)}{N_d} \sum_{j \in \text{different-source}} \log_2 \left( 1 + LR_j \cdot \frac{Pr(\theta_p)}{Pr(\theta_d)} \right) \quad (9)$$

The prior probabilities  $P(\theta_p)$  and  $P(\theta_d)$  are not generally known in forensic evaluation of the evidence, because they depend on many other information sources but the evidence (witnesses, police investigations, other evidences, etc.). Therefore, ECE cannot be computed if prior probabilities are not known. We adopt the solution of representing its value as a prior-dependent magnitude. That leads to the so-called ECE plots, which can be seen in Figure 4. As it is shown, in an ECE plot three performance curves are represented together:

1. The solid curve is the ECE (average information loss) of the LR values computed by the evidence evaluation method under assessment. This is the value which represents the overall performance of a set of LR values, and the lower its value, the better. The interpretation of ECE in terms of information theory is as follows: the higher this ECE curve, the higher the information needed in order to know the true hypotheses on average over cases, and therefore the worse the method. This is the curve obtained from Equation 9, where the x-axis is the logarithm of the prior odds, namely prior log-odds or  $\log_{10} \frac{Pr(\theta_p)}{Pr(\theta_d)}$ .
2. The dashed curve represents the so-called *calibrated* performance. This is obtained by transforming the LR values in the experimental set in

order to *calibrate* them. This process optimises ECE of the original set of LR values preserving the discriminating power of such a set. In other words, the calibrated set of LR values is better than any other set of LR values having the same discriminating power. Therefore, the calibrated performance is always a performance bound, where LR values are not only encouraged to have a high discriminating power, but also to be *calibrated*. According to [16] and [17], this is always a desirable characteristic of a LR in order to lead to *good* posterior probabilities. The calibration transformation can be achieved by the use of the Pool Adjacent Violators algorithm (PAV), described in [17]. The difference among the calibrated performance (dashed curve) and the performance of the LR set (ECE, solid curve) indicates the loss of performance due to calibration problems: the bigger such difference, the less calibrated the method under analysis will be, indicating a calibration problem. This allows the detection of problems that can be solved by the selection of better LR models. It is worth noting that if a fact finder wants to take a decision with the posterior probabilities  $Pr(\theta_p | E, I) = 1 - Pr(\theta_d | E, I)$  according to Equation 2, the optimal decision rule does not change after the calibration transformation. In fact, the role of calibration is improving the performance of the decisions taken with the posterior probabilities in Equation 2, which in this work is measured by ECE. In this sense, the calibrated performance (dashed curve) will be always better (lower value) than the performance of the LR set (solid curve). Details about calibration and the role of the PAV algorithm can be found in [16, 17, 18]. As ECE is decomposed among a discriminating power component and a calibration component [16], the calibrated performance (dashed curve), which minimizes the calibration component without altering the discriminating power, can be viewed as a numerical measure of the discriminating power of the experimental set of LR values. Thus, the lower the dashed curve, the better the discriminating power. In the limit, when there is total separation among same-source and different-source LR values, the calibrated performance will be zero for all values of the prior probability.

3. The dotted curve represents the performance of a method always delivering  $LR = 1$ , referred to as a *neutral* method. This performance is achieved when the evidence gives no information about whether  $\theta_p$  or  $\theta_d$  is true. The posterior probability in this case is equal to the prior prob-

ability, which is independent of the LR experimental set, and therefore the performance of the neutral method is always the same. If the ECE curve of the method under analysis presents a value greater than the curve of neutral performance, then the method will loss more information on average than basing the decisions only on the prior information, *i.e.*, not using the evidence at all. In the range of prior probabilities where this happens, the method at hand should not be used for evidence analysis.

As a summarizing example, the information about the evidence evaluation method can be seen in the ECE curve as follows. For instance, the value of ECE in Figure 4 for a value of  $\log_{10} \frac{Pr(\theta_p)}{Pr(\theta_d)} = 0$  (where  $Pr(\theta_p) = Pr(\theta_d)$ ) is 0.73. This means that the loss of information is reduced by the evidence evaluation method from not knowing the evidence, given by the neutral method (dotted curve) and whose value is 1, to 0.73 once the evidence is analyzed.

## 2.6. Software

Calculation of the LR values were conducted by code written by the authors in R and Matlab. Matlab code for drawing ECE plots can be downloaded from <http://arantxa.ii.uam.es/~dramos/>, based on the FoCal software toolkit<sup>4</sup>.

## 3. Results and Discussion

### 3.1. Descriptive statistics

Descriptive statistics of the observed elements in the considered glass databases (containers (*p*), car and building windows (*w*), Polish building windows (*pl*) and British building windows (*uk*)) are presented in the form of box-plots (Fig. 5). Differences observed in Fig. 5 between the separate use categories (*p* and *w*), or geographical zones (*pl* and *uk*), could be observed (*e.g.* see Fig.5g), however any strong conclusions about origin could not be made on the basis of visual inspection.

---

<sup>4</sup><http://niko.brummer.googlepages.com/focal>

### 3.2. Likelihood ratio calculations - experimental set-up

The following experiments were performed to evaluate the variation of physicochemical properties between objects from various glass categories, or glass objects collected in various countries, on the calculated LR values for pairwise comparisons:

1. Tuning the LR model with data obtained for glass samples from one category, for example window glass ( $w$ ), and generate the LR values comparing results obtained for glass samples from different categories such as container glass ( $p$ ). These experiments were notated to indicate the glasses under consideration, for instance  $w-p$  in this example. In general, the following pairs of experiments were performed  $w-p$ ,  $p-w$ ,  $uk-pl$ ,  $pl-uk$  for SEM-EDX data. These experiments, where the population database and the testing database were different, were referred to as *mismatching* experiments.
2. Tuning the LR model with data obtained for glass samples from one category, for example window glass ( $w$ ) and generate the LR values for pairwise comparisons for glass samples from the same category. These experiments were notated to indicate the pairings, for example  $w-w$  indicates pairwise comparisons of window glasses with window glasses. Specifically the following pairs of experiments were performed  $w-w$ ,  $p-p$ ,  $uk-uk$ ,  $pl-pl$ . These experiments, where the population database and the testing database were of the same type, were referred to as *matching* experiments.

To compute the same-source LR values ( $\theta_p$  is true), 2 of the 4 vectors of elemental composition values from each object were used as control data, and the remaining 2 vectors of elemental composition were used as recovered data. A number of comparisons equal to the number of objects presented in the test set could be conducted. As for these comparisons  $\theta_p$  was known to be true, the expected likelihood ratio in each case was where  $LR > 1$ , and each answer with  $LR < 1$  was a misleading value. For different-source comparisons ( $\theta_d$  is true) the two first vectors of elemental composition for each object were used as control and recovered data respectively. The number of comparisons is therefore the combination, without replacement, of the total number of objects  $m$ , *i.e.*  $(m \cdot (m - 1))/2$ . As for these experiments it was known that  $\theta_d$  was true, the expected likelihood ratio was where the  $LR < 1$ . Each likelihood ratio, where  $LR > 1$ , misleadingly supported  $\theta_p$ .

In the case of matching experiments, where background and comparison databases are of the same type (a *e.g.*  $p$ - $p$  experiment), a problem arises, since the subset of the database for  $p$  data should be split in two parts, one part for background modelling, and a second part for comparison. If the available data is scarce, then small sample size effects are more likely to appear. As a consequence, in this paper a jackknife procedure was employed to use the available data efficiently. For each comparison between two sets of measurements, the rest of measurements in the corresponding subset of the database was used for model tuning. This guarantees that, at each comparison, models would be built with all the available data.

Results of performed experiments are presented in the form of Tippett plots, which includes rates of misleading evidence, and also in the form of ECE plots. Both techniques have been described in Section 2.5.

For both sets of experiments, namely variation of glass procedence (Poland, United Kingdom) and variation of glass type (containers, windows), the SEM-EDX elemental data obtained as a result of the measurements has been modelled following three evidence evaluation approaches:

- **Graphical models**, described in Section 2.3.1. For this work, two different models have been used, namely: *i)* *Standardized* graphical model (Equation 4); and *ii)* Graphical model used in [6] (Equation 5). These models have the advantage of using all the available information from each of the observed elements, which is expected to yield a higher discriminating power than lower-dimensional models. However, they may also require a higher amount of training data than univariate, bivariate or trivariate models.
- **Univariate models**. A likelihood ratio as explained in Section 2.3 is obtained separately for each of the 7 variables involved in the experiments, namely  $Al'$ ,  $Ca'$ ,  $Fe'$ ,  $K'$ ,  $Mg'$ ,  $Na'$  and  $Si'$ . It is expected that each of this separate elements will give moderate discriminating power, but the need for data to train the model is assumed to be lower than for other higher-dimensional problems. Moreover, the individual behavior of each variable can be observed, which may shed some light on the behavior of the other higher dimensional models.
- **Bivariate and trivariate components of graphical models**. Here, the likelihood ratios are computed using the same model as for univariate data (Section 2.3), but using bivariate or trivariate sets of vari-

ables. We have analyzed the multivariate sets of variables which form the components of the graphical models in Equations 4 and 5, namely  $(Na', Si, Ca')$ ,  $(Al', K')$ ,  $(Na', Si')$ ,  $(K', Na')$ ,  $(Al', Fe')$ ,  $(Ca', Na')$  and  $(Ca', Mg')$ . The objective here is two-fold: first, we want to identify combinations of variables which give especially good performance, and on the other hand we want to detect problems of calibration in subsets of variables which, together, seem to work poorly. These results are expected to be better than univariate results in discriminating power, but also to present some sensitivity to the lack of training data.

### 3.3. Results for variation of glass precedence (Poland, United Kingdom)

#### 3.3.1. Graphical models

Results for the standardized graphical model are shown in Figures 6 and 7. It can be seen (Fig. 6a) that the use of a matching background generally lowers the value of ECE, which implies increasing performance. This improvement is especially significant for *uk* data, *uk-uk* being the only case where the ECE curve is lower than the neutral dotted curve (for prior log-odds above  $-0.9$ ). This means that only in the case of *uk - uk* data the use of the evidence evaluation method constitutes a real improvement of performance with respect not using the evidence. The explanation of this bad behavior is given in the remainder of this section. The *uk - uk* case is also the best in terms of discriminating power (because it presents a lower dashed curve), indicating that a LR model for solving comparison problem of *uk* building glass tuned on the *uk* database (matching experiment) have higher ability to extract discrimination information from the evidence than the model tuned on *pl* database. In the *pl* database the adequacy of a matching background is not so clear, but it can be observed a reduction on the maximum values of ECE, which means an improvement of overall performance. The consequences of the effects seen in ECE plots can be also seen in Tippett plots in Figure 7a, where the rates of misleading evidence are much more balanced when the background and test databases are the same (matching experiments).

ECE plots for the graphical model used in [6] are shown in Figure 6b, and its corresponding Tippett plots shown in Figure 7b. In this case, and similarly than for the standardized graphical model, it is clearly seen that there is a significant improvement on the rates of misleading evidence when the database used for background modelling and testing are of the same precedence (*pl - pl* and *uk - uk* cases), *i.e.* see values of false negative and

false positive error rates (respectively SS and DS values) in Figure 7b. In fact, the trend on ECE plots (Fig. 6b) is the same like in the case of a standardised graphical model in Fig 6a. An exception are dashed curves in the case when Polish float glass were used as background and British float glass as a testing data were higher than for results obtained with application of the standardised graphical model (compare third ECE plots in Figs. 6a and b). This means that a LR model supported by application of a standardised graphical model for solving comparison problem of building glass tuned on the Polish database have higher ability to extract discrimination information than a LR model supported by graphical model proposed in [6].

The ECE curves obtained by LR models supported by a standardised graphical model and a LR model supported by graphical model proposed in [6] are far from perfect, and they also present high calibration loss. This is may be due to two main reasons. First, any of the variables involved in the graphical model factorisations (Equations 4 and 5) may behave badly when used for evidence evaluation with the current database, and therefore its influence may transfer to the performance of the whole graphical model. Second, the small size of the databases used in this work may be a problem for the graphical models, which need some amount of data to work properly. In order to confirm such hypotheses, the analysis of the univariate, bivariate and trivariate experiments is motivated, whose results are presented below.

### 3.3.2. Univariate models

We analysed the performance of seven variables of the model separately. For this analysis, we have used each of the variables as univariate data, following the model described in Section 2.3. Example of results in terms of ECE for  $Na'$ ,  $Si'$ ,  $K'$  and  $Fe'$  variables are shown in Figure 8. It can be seen that the ECE plots behave generally better for the matching experiments ( $pl - pl$  in the left column and  $uk - uk$  in the right column) than for the mismatching experiments where the background data is changed (from  $pl - pl$  to  $pl - uk$  or from  $uk - uk$  to  $uk - pl$ ) for variables  $Na'$  and  $Si'$  (e.g. Figs. 8a and b), although for the latter the  $ul - uk$  and  $pl - uk$  experiments behave similarly. That means that the adequacy of a database procedence (Poland or the United Kingdom) affects the performance of such variables. In particular, for those variables the ECE value (solid curve) is lower for matching experiments, which indicates that the overall performance is better. Moreover, it seems that the discriminating power (dashed curve) is also better for matching experiments in the majority of cases. Calibration, which

measures the probabilistic interpretability of LR values, and which can be seen as the difference among the solid and the dashed curves, is also good for some given variables, especially for  $Na'$  and  $Si'$ . This result could be expected, since the descriptive statistics (Section 3.1) showed not a significant difference among univariate data histograms for those variables. Moreover,  $Na'$  (Fig. 8a) and  $Si'$  (Fig. 8b) also exhibit a robust performance when the background database changes, because the ECE values do not significantly change. This makes them especially useful to work in conditions where the degree of matching among the background and testing databases is not known. However, for some variables the effect of the mismatch in the background database is not clearly degrading performance. This happens to  $K'$  (Fig. 8c) and  $Fe'$  (Fig. 8d). Thus, for those variables selecting the background data according to the procedence is not improving performance. It is worth noting, that calibration performance is especially bad for some variables, particularly  $Fe'$  variable (it shows a big difference among the solid and the dashed curves). In fact, from Figure 5c and 5g it can be seen that the behavior of such variables is very incoherent among  $pl$  and  $uk$  databases, and this sensitivity causes the problems when  $pl$  is used to train  $uk$  or vice-versa, because the means and covariances computed in the LR model from training data are not representing the behavior of the test data. Moreover, the scarcity of the measurements for the SEM-EDX database (only 4 measurements are available per object) makes the model much more sensitive to outliers, and also affects the reliability of the within-source distributions assigned in section 2.3. In summary, ECE reveals that  $Na'$  and  $Si'$  variables present a good calibration performance, which allows a good interpretation of the LR as a probabilistic degree of support, and they seem robust to the procedence of the background set. However, the discriminating power of such variables is still poor (the dashed curve is high, and near the dotted curve), which means that such variables are not so useful for evidence evaluation by itself. The rest of variables present high calibration losses and poor performance in terms of ECE, and/or are sensitive to the procedence of the background set.

The rest of variables have been also analyzed, but their results have been omitted for extension. Such analysis yields that the conclusions applied above to  $Na'$  and  $Si'$  are also suitable to  $Ca'$ ; and the conclusions for  $K'$  and  $Fe'$  also apply to  $Al'$  and  $Mg'$ .



### 3.3.3. Bivariate and trivariate components of the graphical models

For this analysis, we have used each of the variables combined according to the components of the proposed graphical models (Equations 4 and 5), following the model described in Section 2.3. Example results in terms of ECE are shown in Figure 9. It can be seen that the models that use variables which behave correctly in terms of calibration in the univariate problems (namely  $Na'$ ,  $Si'$  and  $Ca'$ ) behave better in terms of calibration than the combination of other variables, and its behavior is also quite robust to the selection of the precedence of the glass database. This is the case of the set of variables  $(Na', Si', Ca')$  (Fig. 9(a)), or also the sets of variables  $(Ca', Na')$  and  $(Na', Si')$ , whose results have not been included here for extension of the article. However, although the calibration performance of the model (separatin among solid and dashed curves) is good in most of the cases, there are some cases where it is still sub-optimal (e.g., Figure 13(a), "w w" case). Therefore, we recommend the measurement of ECE performance when such a model is to be used with a given dataset, in order to assess whether the method is appropriate to the database in use. Worth noting,  $(Na', Si', Ca')$  offers better discriminating power (because the dashed curve is lower) than the rest of bivariate set of variables, which suggests that the use of more variables is convenient unless calibration is seriously affected. That motivates the improvements of the calibration of graphical models in Section 3.3.1, which is suggested as future work. It also can be seen that the use of sets of variables containing only variables which performed bad in terms of calibration (Section 3.3.2), and not any of  $Na'$ ,  $Si'$  and  $Ca'$  yields poor performance (ECE value, solid curve), such as for  $(Al', Fe')$  (Fig. 9b) and  $(Al', K')$ , whose results have not been included here for extension of the article. It can be concluded that the combination of such bad results using graphical models yields to the poor performance observed.

In order to highlight the conclusions drawn for the  $(Na', Si', Ca')$  model, Figure 10 shows its performance in terms of Tippett plots. It is seen that the discriminating power of the method is good, with a good ratio among the false-positive and false-negative rates for all the values of the  $x$  axis. However, there are still some effects which negatively affect calibration performance, such as a small but non-negligible amount of comparisons yielding strong misleading evidence (see *e.g.* the *pl-uk* case). These effects are still related to the sparsity of the database, because a trivariate model is still demanding in terms of training data. This sparsity also makes the models more sensitive

to outliers. The study of such effects are outwith the scope of this article, and are proposed as future work.

### 3.4. Results for variation in the glass type (containers, windows)

#### 3.4.1. Graphical models

For the standardized graphical model, results in the form of ECE plots are shown in Figures 11 and 12. The conclusions are similar as in the case of  $pl$  and  $uk$  databases (Section 3.3.1): a matching database improves performance in the case of  $p$  background database, and it is essential in order to have some good calibration. Database mismatch in that case makes calibration unacceptable. On the other hand, for the  $w$  background database the improvement of a matching testing database is not so clear, and performance for both cases is comparable. This can be verified in Tippett plots in Figure 12a, for which an improvement is clearly seen in matching conditions for  $p$  background, and it is not so clear with  $w$  background. The same interpretation can be drawn from results with ECE plots for the graphical model used in [6], shown in Figure 11b, and its corresponding Tippett plots shown in Figure 12a. As a conclusion, matching database conditions is always desirable with SEM-EDX multivariate data. If a matching database is not used the calibration of the evidence evaluation methods can be seriously degraded in some cases.

However, and again as it happened with  $pl - uk$  data, the calibration performance for the  $p - w$  database is poor. In order to obtain an explanation for that fact, we will again analyze the behavior of univariate, bivariate and trivariate problems using ECE plots.

#### 3.4.2. Univariate models

For univariate experiments in the  $p - w$  database, similar conclusions can be extracted than when the  $pl - uk$  database was used (Section 3.3.2). Again, results from variables  $Al'$ ,  $Ca'$  and  $Fe'$  are described but their ECE plots are not included in the article in order to limit its extension. Examples of results in terms of ECE are shown in Figure 13. First, the variables which seem best calibrated (solid curve is close to the dashed curve) are again  $Na'$  (Fig. 13a),  $Si'$  (Fig. 13b) and  $Ca'$ . This means that their interpretability in terms of probabilistic weight of the evidence will be more appropriate than for the rest of variables, and also that they are more suitable for combination than other less-calibrated variables. Moreover, the performance of  $Na'$  variable (ECE, solid curve) is the best of all, which means that it is not only suitable

for combination in a reliable fashion, but also will probably contribute more significantly to the overall performance. Again, these variables seems also more robust to the variation of the type of glass used for training the models, since the variation of the ECE when the background database changes from  $p$  to  $w$  is small. It is also observed that, as it happened in the  $pl - uk$  experimental set-up, the worst variables (ECE, solid curve) are  $Mg'$ ,  $Al'$ ,  $K'$  (Fig. 13c) and  $Fe'$  (Fig. 13d), being set also badly calibrated, and also sensitive to the type of glass used as background. This makes such variables not so reliable for combination using any scheme, *e.g.* graphical models.

### 3.4.3. Bivariate and Trivariate components of the graphical models

Analysis of the variables combined according to the components of the proposed graphical models (Equations 4 and 5) yields similar conclusions as it happened with the precedence of the data ( $pl - uk$  database). Again, results from sets of variables  $(Ca', Na')$ ,  $(Na', Si')$  and  $(Al', K')$  are described but their ECE plots are not included in the article in order to limit its extension. Examples of results in terms of ECE for all the variables are shown in Figure 14. First, we see again that the models including the variables which gave good calibration (namely  $Na'$ ,  $Si'$  and  $Ca'$ ) perform better than the rest. It is the case of the sets of variables  $(Na', Si', Ca')$  (Fig. 14a),  $(Ca', Na')$  and  $(Na', Si')$ . Also,  $(Na', Si', Ca')$  achieves remarkable discriminating power (because its dashed curve is remarkably low), much better than the rest of bivariate set of variables, suggesting again that the use of more variables is convenient unless calibration is seriously affected, and motivating the improvement of the calibration of graphical models in Section 3.3.1, which is suggested as future work. However, although the results are satisfactory and robust when  $w$  is used as background data, results when  $p$  is used as background offer much worse performance (ECE, solid curve), and the calibration seriously degrades (difference among solid and dashed curves). This may be due to outliers, which seriously affects the database used for LR computation, because only 4 measurements are available for each object in the database. This effect is more important in the trivariate  $(Na', Si', Ca')$  set of data, because the more the variables involved, the more the sensitivity to outliers under data scarcity conditions. Again, we also observe that the use of sets of variables containing only variables which performed bad in terms of calibration (Section 3.4.2), such as for  $(Al', K')$  and  $(Al', Fe')$  (Fig. 14b), gives again bad performance (ECE, solid curve), and that these components are the main responsible of the bad behavior of the graphical

models in Section 3.4.1.

As we did for the case of databases of different procedence,  $(Na', Si', Ca')$  models is further analyzed in Figure 15 in terms of Tippett plots. The conclusions for the databases of different glass sizes are equivalent to those of databases of different procedence. First, discriminating power is good. Second, negative effects on the calibration performance still remain, such as the high value of strong misleading evidence in some cases (such as in  $p-w$ ). Again, database sparsity is identified as an important reason of the sub-optimal performance of the model, making it also sensitive to outliers. Future work includes the exploration of such effects in depth.

#### 4. Conclusions

This paper shows results which illustrate the impact of the variation of elemental composition of glass in the performance of the evidence evaluation process, by the generation of experimental results using a database of chemical data in order to calculate LR results. In this database, two main sources of variability have been studied, mainly the type of glass (containers, building and car windows, which divides the database among  $p$  and  $w$  data) and its procedence (Poland, United Kingdom, which divides the database among  $pl$  and  $uk$  data). We have used several models in order to conduct the evidence evaluation process, most of them based on the direct application of the multivariate LR approach, and some of them by the use of graphical models. The performance of the evidence evaluation process has been presented mainly in the form of classical Tippett plots, and also as Empirical Cross-Entropy (ECE) plots, which have an appealing information-theoretical interpretation to represent the discriminating power and calibration performance of the evidence evaluation process.

These results show that the performance of the evidence evaluation process suffers a significant impact in its performance if the population data is not properly selected, *e.g.* if a LR value is computed for control and recovered samples of  $p$  data using the models tuned with  $w$  background data, the performance of the LR values will tend to decrease, especially in terms on calibration performance.

The ECE curves obtained for LR models supported by a standardised graphical model and a LR model supported by graphical model proposed in [6] are far from perfect, and they also present high calibration loss. This is may be due to two main reasons. First, variables involved in the graphical

model factorisations behave badly when used for evidence evaluation with the current database (especially  $Mg'$ ,  $Al'$ ,  $K'$  and  $Fe'$  variables), and therefore its influence may transfer to the performance of the whole graphical model. Second, the small size of the databases used in this work may be a problem for the graphical models, which need some undefined quantity of data to work properly. The analysis of the univariate, bivariate and trivariate combination of seven variables allows conclusions to be drawn as to which variables should be used to calculate likelihood ratios, which not only allow the forensic scientist to say which hypothesis is supported by the evidence, but also the strength of such support.

It was found that LR values computed from  $Na'$ ,  $Si'$  and  $Ca'$  variables present a good calibration performance, which allows a good interpretation of the LR as a probabilistic degree of support. They also seem robust to the precedence of the background data. This means that their interpretability in terms of probabilistic weight of the evidence will be more appropriate than for the rest of variables, and also that they are more suitable for combination than other LR values from other different variables. However, the discriminating power of such variables is still poor, which means that such variables are not so useful for evidence evaluation by itself, because their ability to distinguish same-source from different-source comparisons is limited. The combination of these three variables in one single multivariate model allowed to obtain remarkable discriminating power, much better than the rest of bivariate set of variables, suggesting again that the use of more variables is convenient. However, although the calibration performance of the  $(Na', Si', Ca')$  multivariate model is good in most of the cases, there are some cases where it is still sub-optimal. Therefore, we recommend the measurement of ECE performance when such a model is to be used with a given dataset, in order to assess whether the method is appropriate to the database in use.

The relative ability of the  $(Na', Si', Ca')$  clique to provide good discriminatory measures refute the belief that SEM-EDX method cannot be applied to comparison problems of glass fragments for forensic purposes, as SEM-EDX provides information about main elemental components of glass.

## Acknowledgments

The authors wish to thank Mr. Jim Haworth, Key Forensic Services, University of Warwick Science Park, Coventry, UK, for delivery of samples

of British float glass and Dr. David Lucy, Department of Statistics, Lancaster University, UK for helpful comments and language support.

## References

1. J. M. Curran, T. N. Hicks, J. S. Buckleton, *Forensic Interpretation of Glass Evidence*, CRC Press, 2000.
2. G. Zadora, Z. Brozek-Mucha, SEM-EDX - a useful tool for forensic examinations, *Material Chemistry and Physics* 81 (2003) 345–348.
3. T. Hicks, F. M. Sermier, T. Goldmann, A. Brunelle, C. Champod, P. Margot, The classification and discrimination of glass fragments using non destructive energy dispersive X-ray fluorescenc, *Forensic Science International* 137 (2003) 107–118.
4. T. Trejos, J. R. Almirall, Strategies for the analysis of glass fragments by LA-ICP-MS Part 1: micro-homogeneity study of glass and its application to the interpretation of forensic evidence, *Talanta* 67 (2005) 388–395.
5. C. G. G. Aitken, D. Lucy, G. Zadora, J. M. Curran, Evaluation of trace evidence for three-level multivariate data with the use of graphical models, *Computer Statistics and Data Analysis* 50 (2006) 2571–2588.
6. C. G. G. Aitken, G. Zadora, D. Lucy, A two-level model for evidence evaluation, *Journal of Forensic Sciences* 52 (2) (2007) 412–419(8).
7. G. Zadora, Classification of glass fragments based on elemental composition and refractive index, *Journal of Forensic Sciences* 54 (2009) 49–59.
8. G. Zadora, T. Neocleous, C. Aitken, Two-level model for evidence evaluation in the presence of zeros, *Journal of Forensic Sciences* 55 (2) (2010) 371–384.
9. G. Zadora, T. Neocleous, Evidential Value of Physicochemical Data - Comparison of Methods of Glass Database Creation, *Journal of Chemometrics* 24 (2) (2010). doi: 10.1002/cem.1276.
10. C. G. G. Aitken, F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley & Sons, Chichester, 2004.

11. A. W. F. Edwards, *Likelihood*, The Johns Hopkins University Press, Baltimore, MD, USA and London, UK, 1992.
12. I. W. Evett, Towards a uniform framework for reporting opinions in forensic science casework, *Science and Justice* 38 (3) (1998) 198–202.
13. C. G. G. Aitken, D. Lucy, Evaluation of trace evidence in the form of multivariate data, *Applied Statistics* 53 (2004) 109–122, With corrigendum 665–666.
14. J. Zieba-Palus, G. Zadora, J. Milczarek, Differentiation and evaluation of evidence value of styrene acrylic urethane topcoat car paints analysed by pyrolysis-gas chromatography, *Journal of Chromatography A* 1179 (2008) 47–58.
15. I. W. Evett, J. Buckleton, Statistical analysis of STR data, *Advances in Forensic Haemogenetics*, Springer-Verlag, Heilderberg 6 (1996) 79–86.
16. M. H. deGroot, S. E. Fienberg, The comparison and evaluation of forecasters, *The Statistician* 32 (1982) 12–22.
17. N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, *Computer Speech and Language* 20 (2-3) (2006) 230–275.
18. D. Ramos, Forensic evaluation of the evidence using automatic speaker recognition systems, Ph.D. thesis, Depto. de Ingenieria Informatica, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Madrid, Spain, available at <http://atvs.ii.uam.es> (2007).
19. D. Ramos, J. Gonzalez-Rodriguez, J. Fierrez, Information-theoretical comparison of evidence evaluation methods for score-based biometric systems, in: *Proceedings of International Conference of Forensic Inference and Statistics*, 2008.
20. D. Ramos, J. Gonzalez-Rodriguez, G. Zadora, J. Zieba-Palus, C. G. G. Aitken, Information-theoretical comparison of likelihood ratio methods of forensic evidence evaluation, in: *Proceedings of International Workshop on Computational Forensics (in IAS 2007)*, 2007, 411–416.
21. B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 1986.

22. J. Whittaker, Graphical models in applied multivariate statistics, John Wiley and Sons, 1990.
23. G. Zadora, T. Neocleous, Likelihood ratio model for classification of forensic evidences, *Analitica Chimica Acta*, 64 (2009) 266-278.
24. T. M. Cover, J. A. Thomas, Elements of Information Theory, 2nd ed., Wiley Interscience, 2006.



## 5. Figures

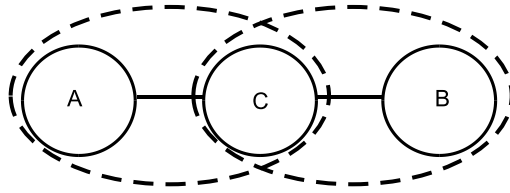


Figure 1: Example of graphical model. Two cliques are marked by dashed ellipses.

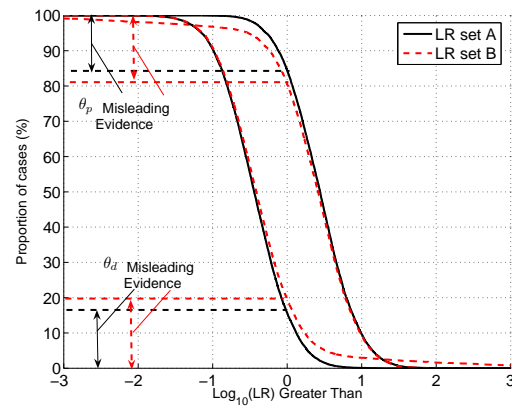


Figure 2: Example of two LR values sets with similar rates of misleading evidence and very different behavior.

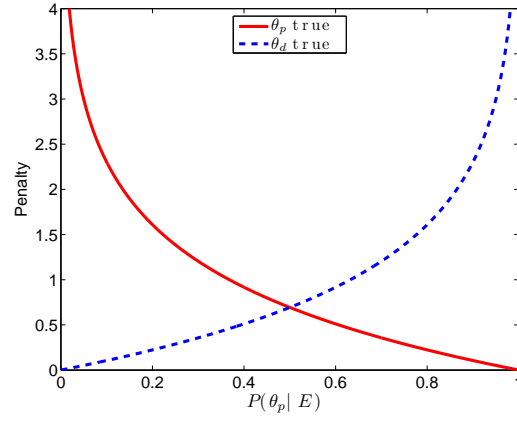


Figure 3: Logarithmic scoring rule. The x-axis represents the posterior probability of  $\theta_p$ , which may be viewed as the “forecast” of about whether  $\theta_p$  is true considering all the available knowledge in a given forensic case, including the evidence  $E$ .

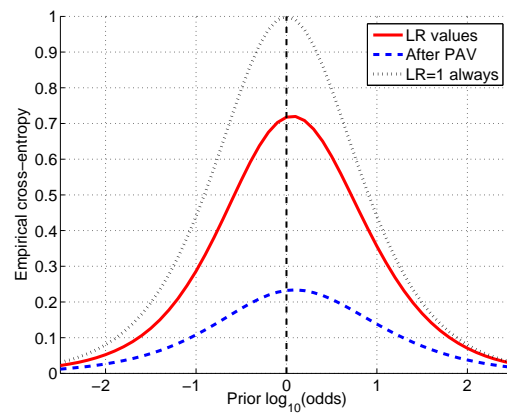


Figure 4: Example of ECE plot.

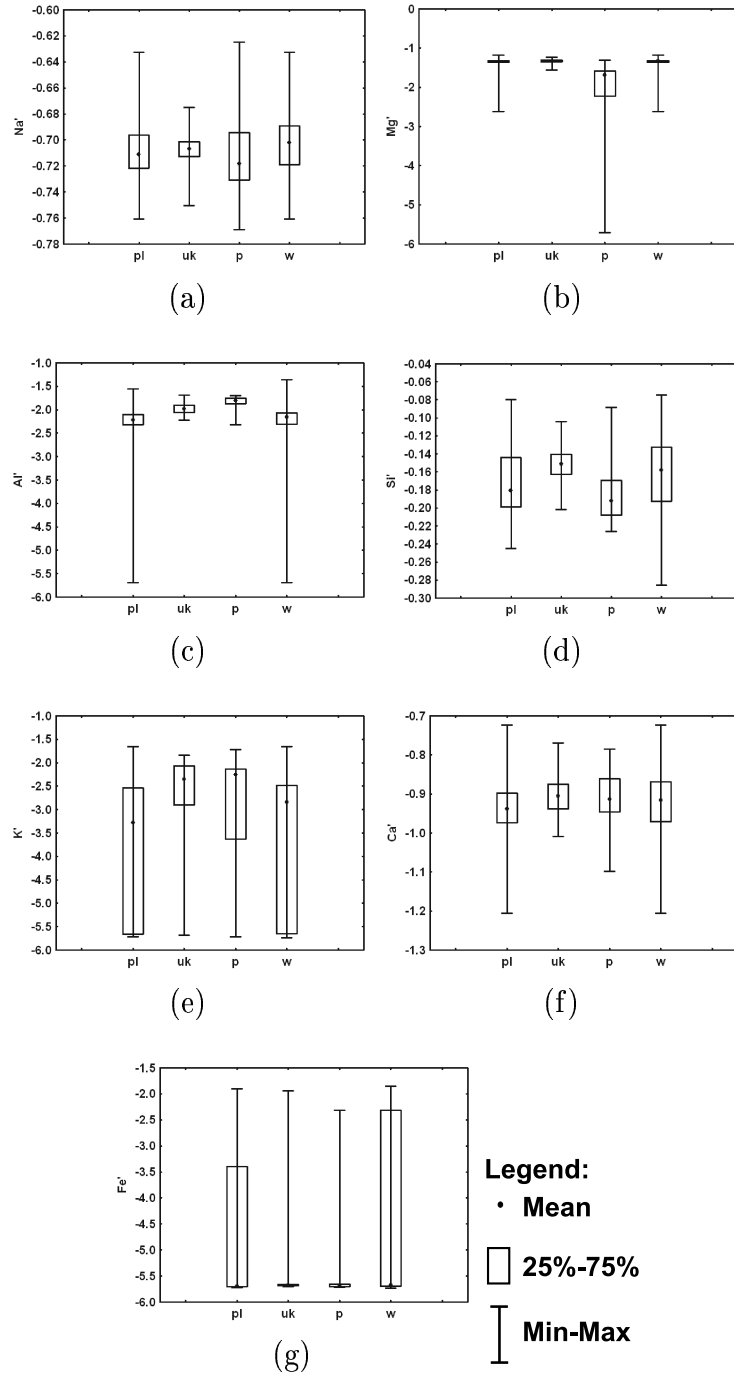


Figure 5: Box-plots - distributions of variables in each of the four glass categories (*p* - container, *w* - car and building windows, *pl* - Polish building windows, *uk* - British building windows).

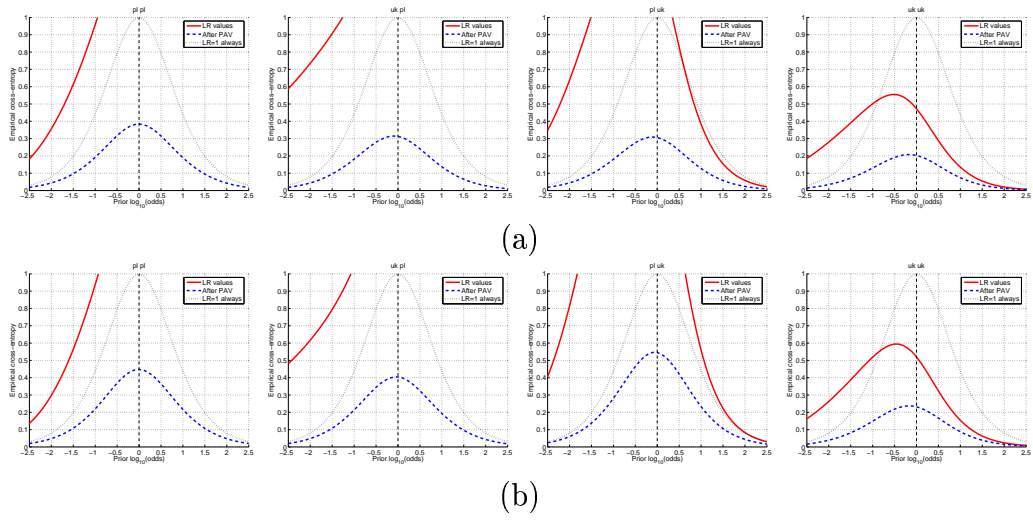


Figure 6: Variation among *pl* and *uk* databases. ECE performance for the SEM-EDX experiment with standardized graphical model (a), and with graphical model used in [6](b). Each column is one experimental configuration, from left to right: *pl* used as background and *pl* as testing data; *uk* used as background and *pl* as testing data; *pl* used as background and *uk* as testing data; *uk* used as background and *uk* as testing data.

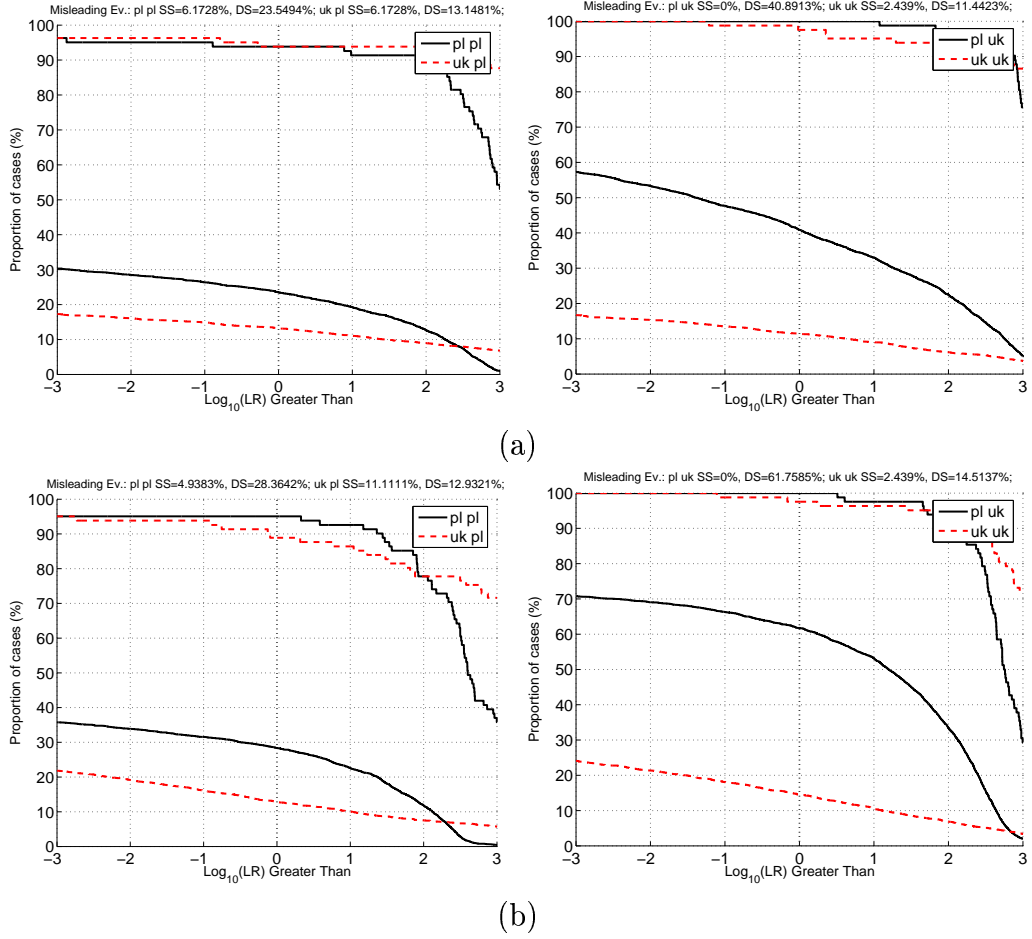


Figure 7: Variation among *pl* and *uk* databases. Tippet plots for the SEM-EDX experiment. (a): standardized graphical model, (b): graphical model used in [6]; from the left to right - *pl* or *uk* used as testing data, varying background data. Error rates at a LR value of 1 are represented in the top of each figure, namely SS (*same-source experiment*) rate of false negative cases; and DS (*different-sources experiment*) rate of false positive cases.



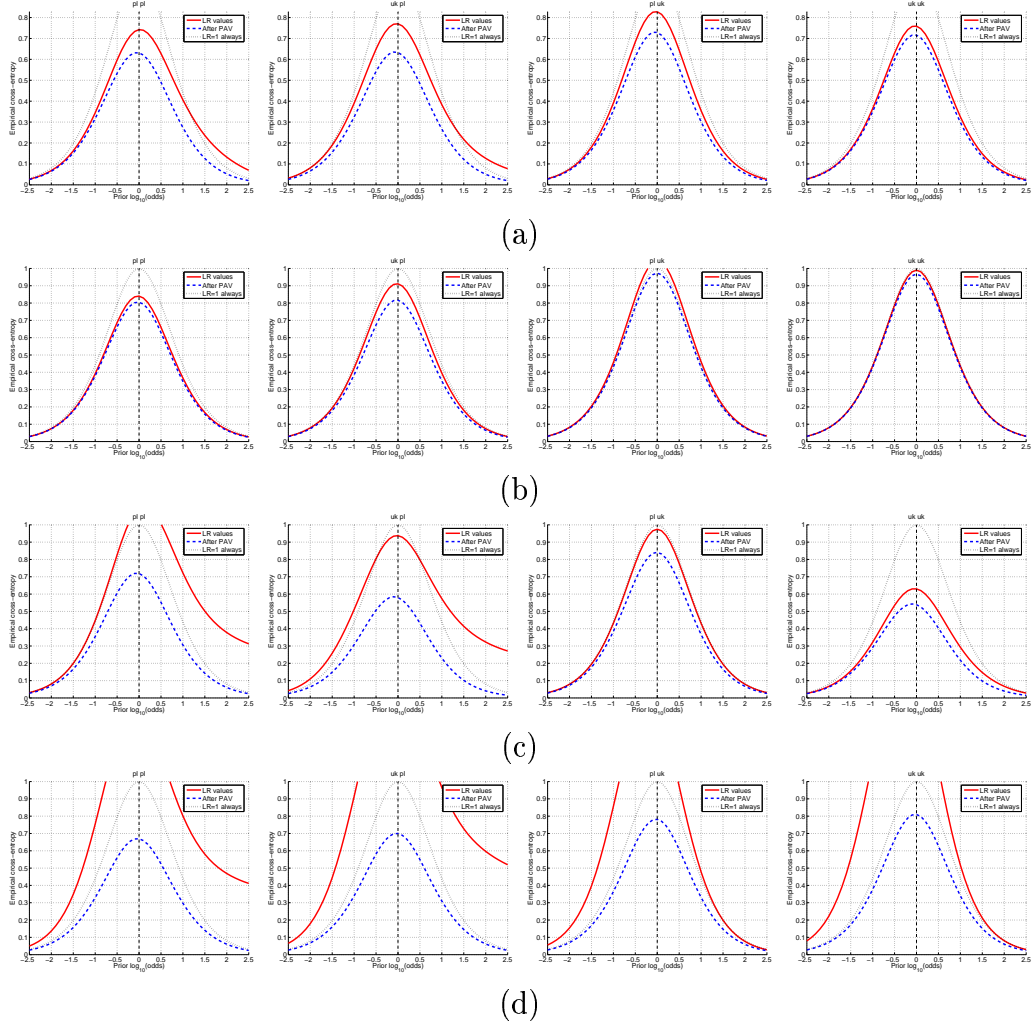


Figure 8: Variation among  $pl$  and  $uk$  databases. ECE performance for the SEM-EDX experiment and for different univariate data, namely  $Na'$  (a),  $Si'$  (b),  $K'$  (c) and  $Fe'$  (d). Each column is one experimental configuration, from left to right:  $pl$  used as background and  $pl$  as testing data;  $uk$  used as background and  $pl$  as testing data;  $pl$  used as background and  $uk$  as testing data;  $uk$  used as background and  $uk$  as testing data.

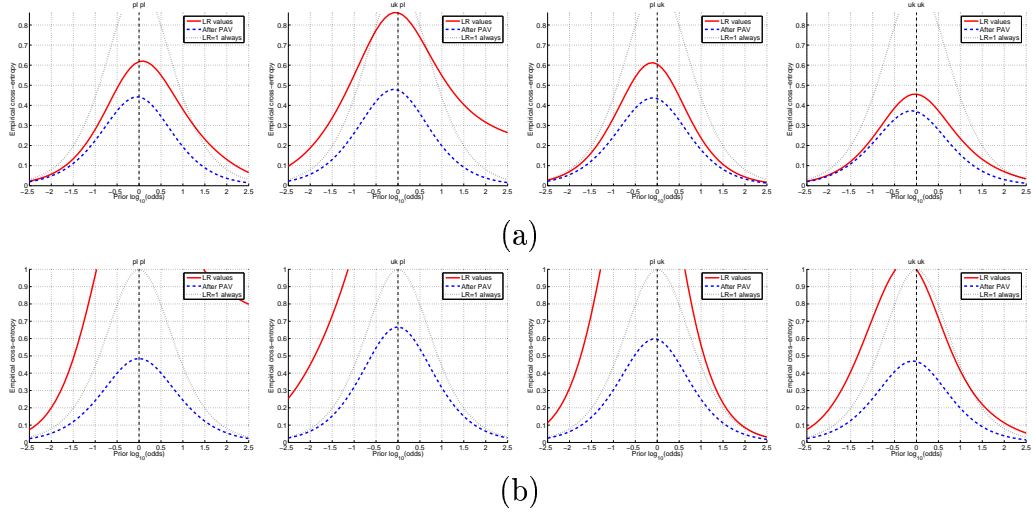


Figure 9: Variation among *pl* and *uk* databases. ECE performance for the SEM-EDX experiment and for different bivariate and trivariate data, namely  $(Na', Si', Ca')$  (a) and  $(Al', Fe')$  (b). Each column is one experimental configuration, from left to right: *pl* used as background and *pl* as testing data; *uk* used as background and *pl* as testing data; *pl* used as background and *uk* as testing data; *uk* used as background and *uk* as testing data.

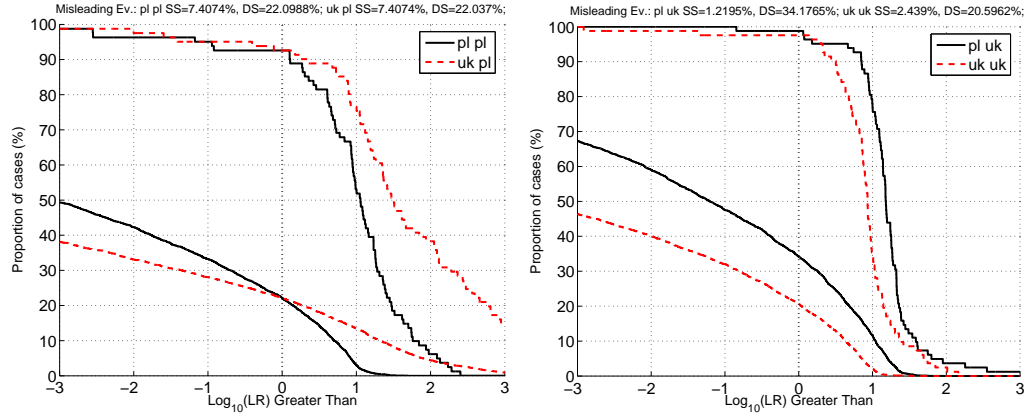


Figure 10: Variation among *pl* and *uk* databases. Tippett plots for the  $(Na', Si', Ca')$  experiment. From the left to right - *pl* or *uk* used as testing data, varying background data. Error rates at a LR value of 1 are represented in the top of each figure, namely SS (*same-source experiment*) rate of false negative cases; and DS (*different-sources experiment*) rate of false positive cases.

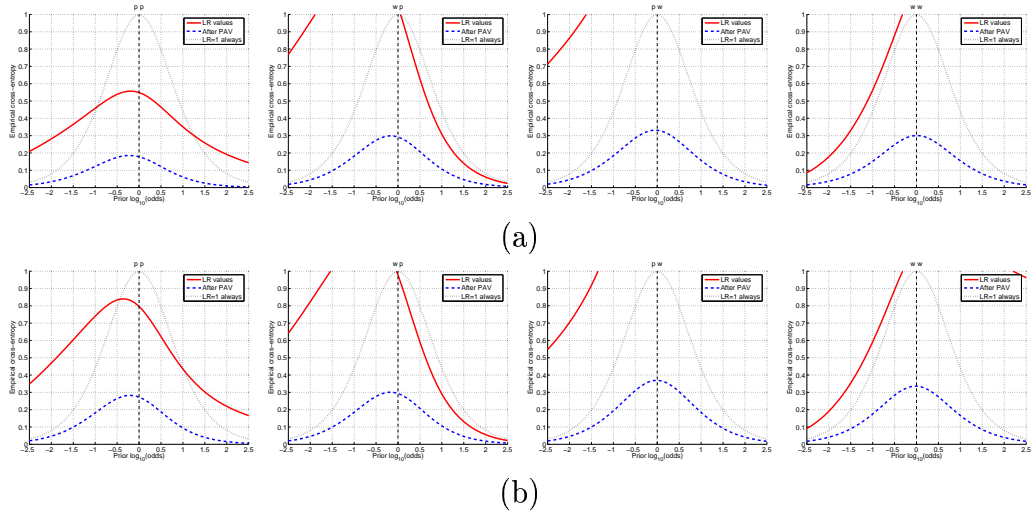


Figure 11: Variation among  $p$  and  $w$  databases. ECE performance for the SEM-EDX experiment with standardized graphical model (a), and with graphical model used in [6](b). Each column is one experimental configuration, from left to right:  $p$  used as background and  $p$  as testing data. (b):  $p$  used as background and  $w$  as testing data. (c):  $w$  used as background and  $p$  as testing data. (d):  $w$  used as background and  $w$  as testing data.

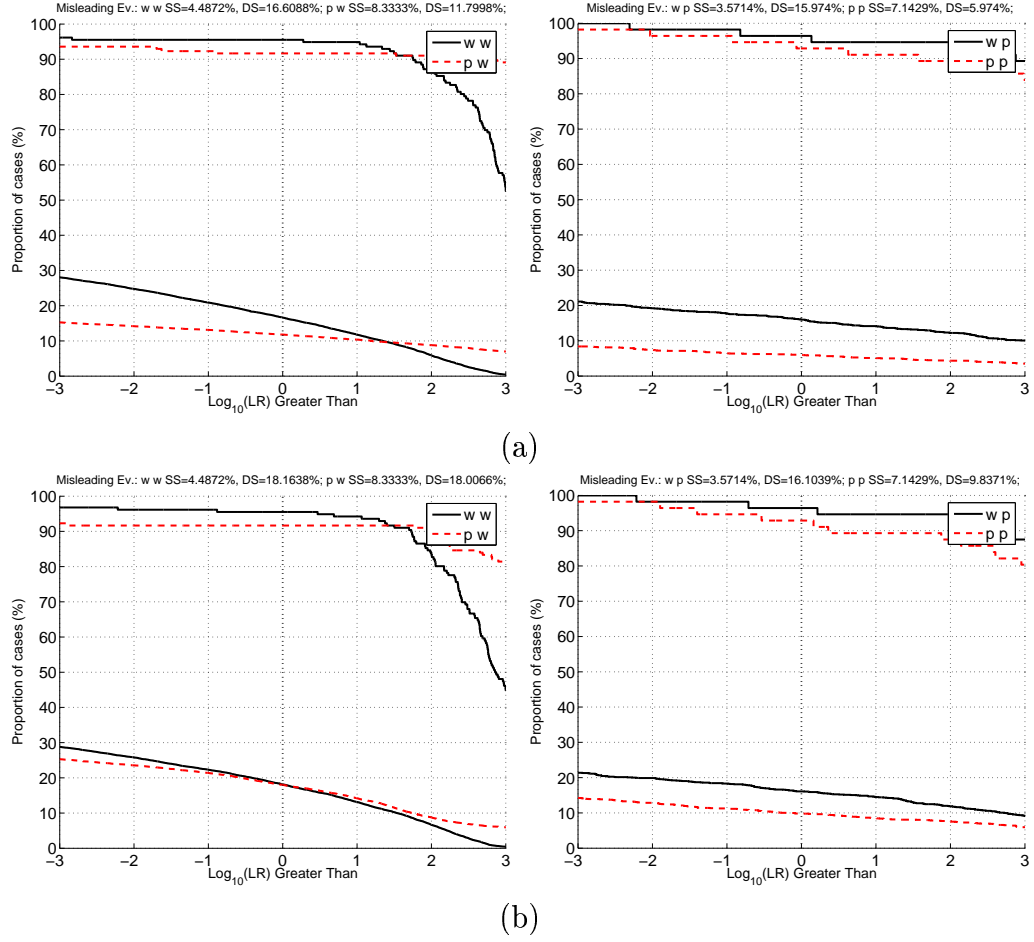


Figure 12: Variation among  $p$  and  $w$  databases. Tippet plots for the SEM-EDX experiment (standardized graphical model). ((a): standardized graphical model, (b): graphical model used in [6]; from the left to right -  $p$  or  $w$  used as testing data, varying background data. Error rates at a LR value of 1 are represented in the top of each figure, namely SS (*same-source experiment*) rate of false negative cases; and DS (*different-sources experiment*) rate of false positive cases.

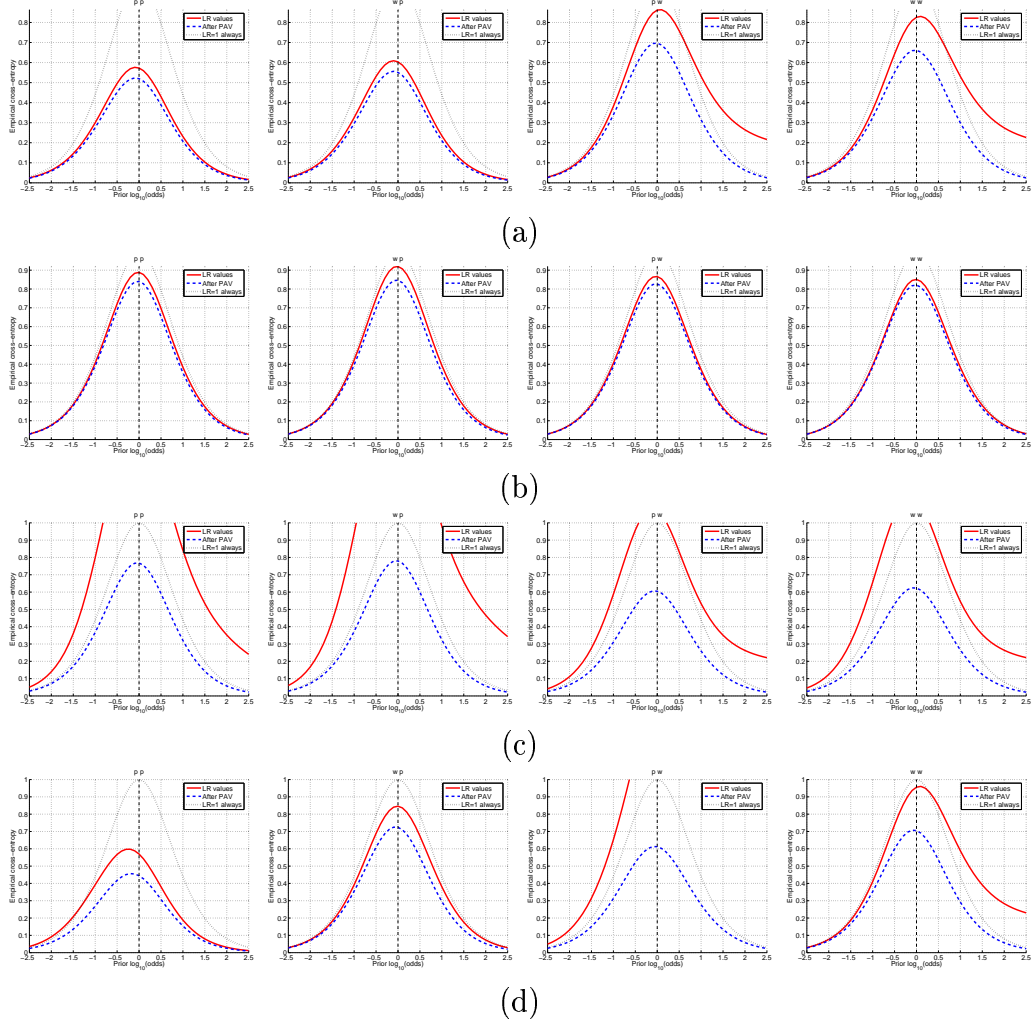


Figure 13: Variation among  $p$  and  $w$  databases. ECE performance for the SEM-EDX experiment and for different univariate data, namely  $Na'$  (a),  $Si'$  (b),  $Fe'$  (c) and  $K'$  (d). Each column is one experimental configuration, from left to right:  $p$  used as background and  $p$  as testing data;  $w$  used as background and  $p$  as testing data;  $p$  used as background and  $w$  as testing data;  $w$  used as background and  $w$  as testing data.

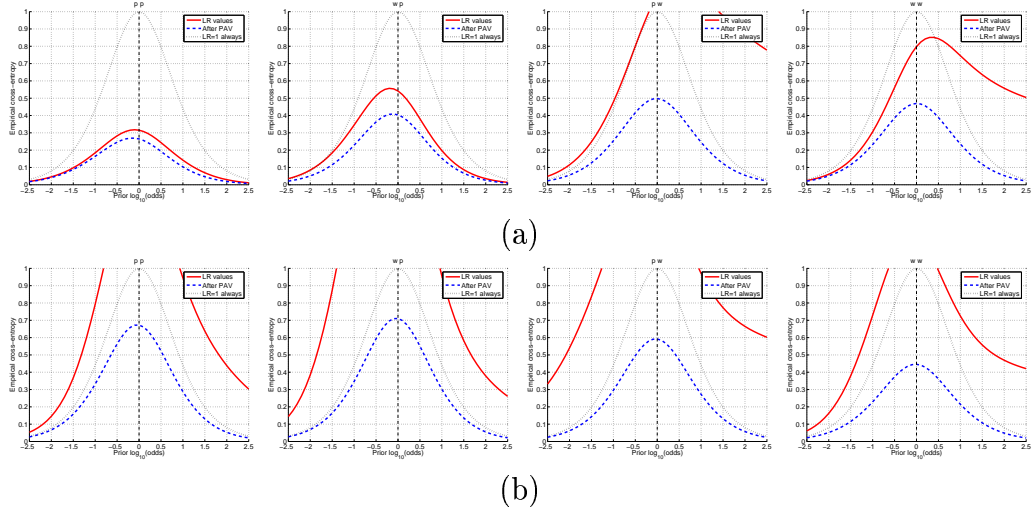


Figure 14: Variation among  $p$  and  $w$  databases. ECE performance for the SEM-EDX experiment and for different bivariate and trivariate data, namely  $(Na', Si', Ca')$  (a) and  $(Al', Fe')$ . Each column is one experimental configuration, from left to right:  $p$  used as background and  $p$  as testing data;  $w$  used as background and  $p$  as testing data;  $p$  used as background and  $w$  as testing data;  $w$  used as background and  $w$  as testing data.

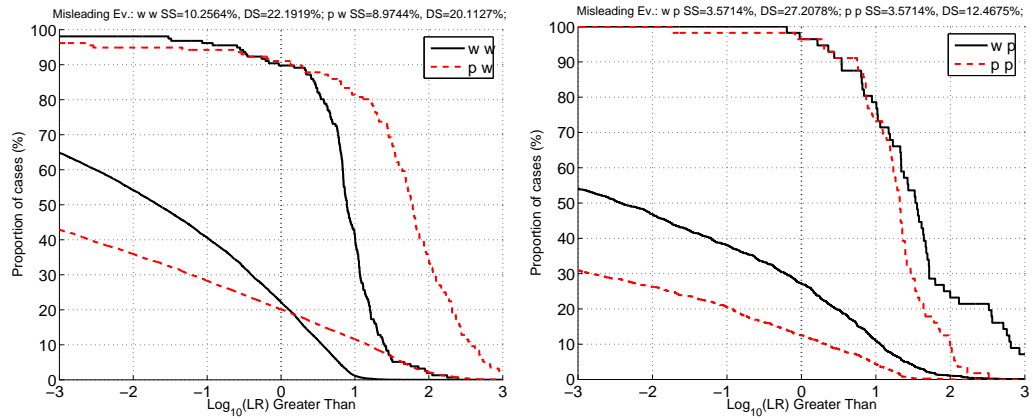


Figure 15: Variation among  $p$  and  $w$  databases. Tippett plots for the  $(Na', Si', Ca')$  experiment. From the left to right -  $w$  or  $p$  databases used as testing data, varying background data. Error rates at a LR value of 1 are represented in the top of each figure, namely SS (*same-source experiment*) rate of false negative cases; and DS (*different-sources experiment*) rate of false positive cases.