Analytica Chimica Acta 705.1-2 (2011): 207 – 217

DOI: http://dx.doi.org/10.1016/j.aca.2011.05.029

Information-theoretical feature selection using SEM-EDX data for the classification of glass traces.

Daniel Ramos[a], Grzegorz Zadora[b]

[a] *ATVS - Biometric Recognition Group, Universidad Autonoma de Madrid, 28049 Madrid, Spain.*
[b] *Institute of Forensic Research, Westerplatte 9, 31-033 Krakow, Poland.*

Abstract

In this work, a selection of the best features for multivariate forensic glass classification using Scanning Electron Microscopy coupled with an Energy Dispersive X-ray spectrometer (SEM-EDX) is performed. This is motivated since the databases available in this forensic glass classification is sparse nowadays, and the acquisition of SEM-EDX data is costly and time-consuming for forensic laboratories. The database used consists of 278 glass objects, for which 7 variables based on elemental compositions of glass obtained with SEM-EDX are available. Two categories are considered for the classification task, namely containers and car/building windows, both of them typical in forensic casework. A multivariate model is proposed for the computation of likelihood ratios. The feature selection process is carried out by means of an exhaustive search, with an Empirical Cross-Entropy (ECE) objective function. The ECE metric takes into account not only the discriminating power of the model in use, but also its calibration, which indicates whether likelihood ratios are interpretable in a probabilistic way. Thus, the proposed model is applied to all the 63 possible univariate, bivariate and trivariate combinations taken from the 7 variables in the database, and its performance is ranked by its ECE. Results show **remarkable accuracy of the best variables selected following the proposed procedure for the task of classifying into windows (from cars or buildings) or containers**, obtaining high (almost perfect) discriminating power and good calibration. This allows the proposed models to be used in casework. We also

present a deep analysis which reveals the benefits of the proposed ECE metric as an assessment tool for classification models based on likelihood ratios.

Key words:
Glass classification, feature selection, Empirical Cross-Entropy, forensic evaluation of the evidence, likelihood ratio, physico-chemical multivariate data, SEM-EDX.

---

1. Introduction

1.1. Glass analysis for forensic purposes

Scanning Electron Microscopy coupled with an Energy Dispersive X-ray spectrometer (SEM-EDX) is routinely used in many forensic institutes for the investigation of glass and other forensic samples [1]. It is because this methods allow to work with very small objects like glass fragments (0.1 - 0.5 mm) that arise during car accidents, burglaries, fights *etc.* This is a strong advantage of this method in contrary to other analytical techniques used for the determination of elemental composition of glass fragments, which require relatively large fragments of glass (larger than 0.5 mm), like μ-X-Ray Fluorescence [2] and Laser Ablation-Inductively Coupled Plasma-Mass Spectrometry [3]. 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 [4–7].

In this work, the problem of glass classification into one amongst different types (windows, containers *etc*) is presented, and a solution based on a multivariate likelihood ratio of elemental concentrations is described based on previous work on glass comparison [8] and classification tasks [6]. The original contribution relies on an exhaustive feature selection process in order to find the best

variables for the task, according to an information-theoretical objective measure of performance. This work is organised as follows. The rest of Section 1 introduces the basics of the likelihood ratio approach for evidence evaluation in glass classification tasks. Section 2 describes in detail the methods and models used in this article. Moreover, the information-theoretical measure of performance used as an objective function for the selection of variables is presented, namely *Empirical Cross-Entropy* (ECE) and its summarizing C_{llr} measure. In Section 3 the exhaustive feature selection process is described, and the results are presented, where the best variables for glass classifications are identified and ranked. Finally, conclusions are drawn in Section 4.

1.2. Likelihood ratio framework for forensic evidence evaluation

Standard modern forensic practice is to interpret physicochemical data as evidence (E), in the context of two competing hypotheses or propositions, one proposed by the prosecutor (θ_p), and one proposed by the defence (θ_p), which is usually complementary[1]. Evidence in forensic science is understood as the relationship of some questioned, *recovered* material; and some *control* material of known origin. In this work, the **features** will be vectors of measurements of elemental composition of glass, but in general they may be DNA profiles, speech features, etc. Such evidence evaluation could be numerically done by applying approaches based in a Bayesian paradigm and related to a likelihood ratio (LR). Bayes' theorem [9] relates conditional probabilities and their transposes:

$$Pr\left(\theta_p \mid E, I\right) = \frac{Pr\left(E \mid \theta_p, I\right) \cdot Pr\left(\theta_p \mid I\right)}{Pr\left(E \mid I\right)} \tag{1}$$

where I is the background information available in the case, and not related to the evidence E. This I may include not only circumstantial information in the case (such as witness testimony or police investigations), but also the analysis of other forensic evidence apart from E (for instance, E may be evidence in the

[1] Although other scenarios are possible, in this work we only consider the case of θ_p and θ_d being complementary.

form of elemental concentrations, but other evidence can be present in the case in the form of refractive index analysis). Equation 1 then allows the following inference:

$$\frac{Pr\left(\theta_p \mid E, I\right)}{Pr\left(\theta_d \mid E, I\right)} = \frac{Pr\left(E \mid \theta_p, I\right)}{Pr\left(E \mid \theta_d, I\right)} \cdot \frac{Pr\left(\theta_p \mid I\right)}{Pr\left(\theta_d \mid I\right)} \tag{2}$$

Equation 2 is the *odds* form of Bayes' theorem, where the term *odds* refers to as the quotient of two complementary probabilities. In a typical situation, a 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. The relevant propositions for the fact finder arise from the circumstances of the case, and often because of the adversarial nature of the system. For a case for which the category of origin of a given glass fragment is to be clarified given the evidence, such hypotheses could be:

- θ_p: the recovered glass fragment come from category A,
- θ_d: the recovered glass fragment come from category B.

The definition of these hypotheses determine a so-called *classification problem*, a typical scenario in forensic science.

The Bayesian inference process implicit in a likelihood ratio framework distinguish two main quantities:

1. The prior probabilities $Pr\left(\theta_p \mid I\right) = 1 - Pr\left(\theta_d \mid I\right)$, which are province of the fact finder, and should be stated assuming only the background information (I) in the case, *i.e.*, not considering the evidence E to be analyzed by the forensic scientist [10]
2. The likelihood ratio $LR = \frac{Pr(E|\theta_p,I)}{Pr(E|\theta_d,I)}$ [2], computed by the forensic scientist [9, 11]. Values of LR above 1 support θ_p and values of LR below 1 support θ_d. A value of LR close to 1 provides little support for either proposition.

[2] The background information I is always conditioning all the probabilities, and will be eliminated form the notation for simplicity.

Also the larger than 1 (the lower than 1) the value of the LR, the stronger the support of the evidence for θ_p (θ_d). Likelihood ratio is considered to be the best documented, and most developed, measure of evidential value for forensic purposes [9].

There is an increasing number of applications of this approach in evaluation of physicochemical data for univariate or multivariate data [7, 12], in particular where the observations of elemental composition were by means of SEM-EDX [4–8] or paint data [13]. This is because it allows to objectively evaluate physicochemical data within a classification process including information about:

1. possible sources of uncertainty (sources of error) which will include, at least:
   (a) variation of measurements of characteristics within the recovered and/or control items,
   (b) variation of measurements of characteristics between various objects in the relevant population (*e.g.* glass objects population),
2. information about the rarity of the determined physico-chemical characteristics (e.g.: elemental and/or chemical composition of compared samples) for recovered and/or control samples in the relevant population;
3. the level of **correlation** between different characteristics when more than one characteristic has been measured.

1.3. Classification problem of glass objects

In this work, we use the models proposed in [14], where glass fragments were classified into different types using data obtained by SEM-EDX and GRIM methods. This model **takes into account variability of the considered variables among classified objects (between-object variability) and also within-object variability. This variability is modelled from a database, assuming a given distribution of the data in the database.** Within this model a multivariate kernel density approach was adopted for modelling between-objects distributions and a multivariate normal distribution was

adopted for modelling within-objects distributions. It worked relatively efficiently, except that the assignment to car windows (c) and building windows (w) needs to be treated with care due to the very similar elemental content of these two categories. This fact is a consequence of the similarities in the way they are manufactured (a float glass manufacturing method). The differentiation between these two categories was possible when information about refractive indices values determined before (RI_b) and after (RI_a) an annealing process (namely $dRI = log10|RIa - RIb|$) was used. Moreover, research in [6] shows some evidence that the performance of the LR model is comparable to some existing classification techniques such as Support Vector Machines (SVM) and Naive Bayes Classifiers (NBC). Although the application of SVM and NBC gave slightly better results in [6] than the application of the LR model in terms of classification error rates, the observed differences were not great, and no single classification method was clearly more effective than any other. However, the clear advantage of the LR model is its probabilistic interpretation in a forensic context. This is of critical relevance, since it allows the integration of the evidence evaluation results in the Bayesian reasoning process (Equation 2)[9, 15]. **The importance of a LR-based evidence evaluation paradigm can be justified in recent works such as [16].**

1.4. Performance of LR models

In forensic science, each method of data evaluation should be treated like a supportive tool. Therefore, their performance should be analysed in detail, in order to determine whether it delivers useful information for the evaluation of the evidential value of the obtained physico-chemical data. In this sense, given Equation 2, the LR value should increase the posterior odds in cases when θ_p is true (the recovered glass is actually from category A), and should decrease the posterior odds in cases where θ_d is true (the recovered glass is actually from category B). As a consequence, for cases where θ_p is true then LR should be greater than 1, and for cases where θ_d is true then the LR should be lower than 1. If a given LR value does not behave in this way, it is said to be yielding

misleading evidence. Therefore, misleading evidence happens when LR>1 in cases when θ_d is true, or when LR<1 in cases when θ_p is true. Moreover, if LR$>>1$ in cases when θ_d is true, or LR$<<1$ in cases when θ_p is true; we will say that the LR values are yielding *strongly misleading* evidence [17]. Obviously, and roughly speaking, the stronger the misleading evidence, the more wrong the decisions taken in the light of the evidence will tend to be, since such decisions will be usually taken from the posterior odds in Equation 2. Therefore, for each case in which a material of glass is to be classified, the obtained value of the LR should not deliver strongly misleading evidence. Thus, it would be desirable that when a LR value yields misleading evidence its value should be rather close to 1, and when the evidence supports the correct hypothesis then LR values should be far from 1. This problem cannot be analysed by simple calculation of false classification rates, since the magnitude of the value of the LR in a falsely classified glass case is not taken in to account by the rate. Recently, an approach for measuring the *goodness* or *performance* of the likelihood ratios obtained in a Bayesian framework has been proposed for forensic evaluation of the evidence. It is based on an information-theoretical magnitude, the Empirical Cross-Entropy (ECE), and its representation in the form of the so-called ECE plots [18], which measure the information introduced by the evidence evaluation in the inference process. It is based on strictly proper scoring rules [19, 20], and has been proposed as a proper way of assessing the performance of LR values in different disciplines in forensic science [21, 22]. It was successfully applied in the analysis of performance of LR models used for solving comparison problems using glass samples [8, 22] and paints [22].

2. Methods, models and data

2.1. Glass database

The database used in this paper consists of 278 glass-objects, which include 79 containers (p), 199 float glass (w, *i.e.* 94 car windows (c)and 105 building windows (w)).

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

In this paper, the likelihood ratio model proposed in [14] was used. It considers two levels of uncertainty for multivariate data, namely within- and between-object variability; and it is based on the following assumptions:

1. Consider a number of m_g glass objects belonging to a given class g, where for this work $g \in \{1, 2\}$. Each object is described by k variables and each variable is measured n times on a continuous scale within each object. Therefore, the data are in the form of $k-$dimensional vectors $\mathbf{x}_{gij} = (x_{gij1}, \ldots, x_{gijk})^T$ for $(i = 1, \ldots, m_g)$ and $(j = 1, \ldots, n)$. The object means of these measurements are $\bar{\mathbf{x}}_{gi} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_{gij}$.

2. 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:

$$\mathbf{U}_g = \frac{\mathbf{S}_{wg}}{m_g n - m_g} \qquad \text{where} \qquad \mathbf{S}_{wg} = \sum_{i=1}^{m_g}\sum_{j=1}^{n}(\mathbf{x}_{gij}-\bar{\mathbf{x}}_{gi})(\mathbf{x}_{gij}-\bar{\mathbf{x}}_{gi})^T.$$

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

$$\mathbf{C}_g = \frac{\mathbf{S}_g}{m_g - 1} - \frac{\mathbf{S}_{wg}}{n(nm_g - m_g)} \qquad \text{where} \qquad \mathbf{S}_g = \sum_{i=1}^{m_g}(\bar{\mathbf{x}}_{gi}-\bar{\mathbf{x}}_{\mathbf{g}})(\bar{\mathbf{x}}_{gi}-\bar{\mathbf{x}}_{\mathbf{g}})^T.$$

4. The between-source distribution is estimated from the group means using a multivariate kernel density function with normal kernel functions having each one mean $\bar{\mathbf{x}}_{gi}$ and covariance $h_g^2\mathbf{C}_g$ [5]. According to [23], an optimal value, for the window smoothing parameter h_g for the kernel distribution is expressed as:

$$h_g = \left(\frac{4}{2k+1}\right)^{\frac{1}{k+4}} \frac{1}{m_g^{\frac{1}{k+4}}}.$$

5. The vectors $\mathbf{y}_j = (y_{j1}, \ldots, y_{jk})^T$ where $(j = 1, \ldots, n)$ are the measurements on the glass object to be classified into categories $g = 1$ or $g = 2$, being its mean value $\bar{\mathbf{y}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{y}_j$.

The likelihood ratio is finally $LR = \frac{Pr(E|\theta_p,I)}{Pr(E|\theta_d,I)}$, where the numerator $(Pr(E|\theta_p))$ (for which θ_p is assumed true) and the denominator $(Pr(E|\theta_d))$ (for which θ_d is assumed true) could be respectively expressed by Equation 3, with $g = 1$ for numerator and $g = 2$ for denominator:

$$(2\pi) \mid \frac{\mathbf{U}_g}{n_c} + h^2\mathbf{C}_g \mid^{-1/2} \frac{1}{m_g}\sum_{i=1}^{m_g} \exp -\frac{1}{2}(\bar{\mathbf{y}} - \bar{\mathbf{x}}_{gi})^T(\frac{\mathbf{U}_g}{n_c} + h^2\mathbf{C}_g)^{-1}(\bar{\mathbf{y}} - \bar{\mathbf{x}}_{gi}) \quad (3)$$

2.4. Empirical Cross-Entropy

In order to solve the drawbacks in the use of classification error rates, and assessment metric based on information theory [24], namely Empirical Cross-Entropy (ECE) [18, 22] is proposed. The motivation of such a measure is derived from the statistics literature. In [19], 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[3]:

$$\begin{aligned} \theta_p \text{ true}: & -\log_2 \left(Pr\left(\theta_p | E\right)\right) \\ \theta_d \text{ true}: & -\log_2 \left(Pr\left(\theta_d | E\right)\right) \end{aligned} \tag{4}$$

Thus, strictly proper scoring rules may be viewed as loss functions which assign a penalty to a given value of a given posterior probability $Pr\left(\theta_p | E\right) = 1 - Pr\left(\theta_d | E\right)$ depending on: i) the value of the posterior probability, and ii) the true hypothesis among θ_p and θ_d which actually occurred [19, 20][4]. For example, if a probabilistic, metheorologic 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 1.

The proposed measure of goodness, namely Empirical Cross-Entropy, is derived as the average, prior weighted, value of the logarithmic scoring rule for a given set of likelihood ratio values from a given experimental set-up. The ECE value is expressed as follows:

[3]**The adequacy of the logarithmic scoring rule with respect to other proper scoring rules has been highlighted in [20].** 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.

[4]See [19, 20] for more examples of strictly proper scoring rules

$$ECE = \frac{Pr\left(\theta_p\right)}{N_p} \sum_{i \in \text{CathegoryA}} \log_2 \left(1 + \frac{1}{LR_i \cdot \frac{Pr(\theta_p)}{Pr(\theta_d)}}\right) + \frac{Pr\left(\theta_d\right)}{N_d} \sum_{j \in \text{CathegoryB}} \log_2 \left(1 + LR_j \cdot \frac{Pr\left(\theta_p\right)}{Pr\left(\theta_d\right)}\right) \quad (5)$$

where Category A and Category B refer to the comparisons in the experimental set of LR values when θ_p and θ_d are respectively true, and N_p and N_d are the number of such respective experiments.

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. **In other workds, the higher the ECE value, the more the uncertainty remains about the true value of the hypotheses.** This information should be interpreted on average over different forensic cases, 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].

Moreover, it is also demonstrated in [19] that ECE can be divided into two components:

1. A *calibration loss* component, which measures how similar are the forecasts to the frequency of occurrence of θ_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. Calibration can be viewed as a measure of the adequacy of the probabilistic interpretation of a likelihood ratio [18, 25]. Therefore, the lower the calibration loss, the more probabilistically interpretable are the likelihood ratio values in the experimental set. As a consequence, calibration is therefore essential for a Bayesian reasoning approach in forensic science [19, 25].
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. Refinement can be viewed as a measure of the discriminating power of the set of LR values [18, 25]. Thus, the lower the refinement loss, the more *separated* will be the LR values when θ_p is true and the LR values when θ_d is true.

2.5. Representing Empirical Cross-Entropy: ECE plots

The prior probabilities $Pr(\theta_p)$ and $Pr(\theta_d)$ 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 then adopt the solution of representing its value as a prior-dependent magnitude was adopted. That leads to the so-called ECE plots, which can be seen in Figure 2. As it is shown, in an ECE plot three performance curves are represented together:

1. The solid, red 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 5, 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, blue curve represents the *calibrated* performance, calculated by transforming the LR values in the experimental set. This optimises the ECE of the original set of LR values whilst preserving refinement (*i.e.*, discriminating power). 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 have to be *calibrated*. The calibration transformation can be conducted using a Pool Adjacent Violators algorithm (PAV), which reduces the calibration loss of a set of LR values to a minimum value, while preserving its refinment. Details about the PAV algorithm can be found in [25]. Thus, the difference between the calibrated performance (dashed curve), and the performance of the LR experimental set (ECE, solid curve), indicates the loss of performance due to calibration problems: the bigger the difference, the less calibrated the method under analysis will be, which indicates a calibration problem. Finally, as ECE is decomposed among a discriminating power component and a calibration component [19], 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 limiting case, when there is total separation among same-source and different-source LR values, the calibrated performance (dashed curve) 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 θ_p or θ_d is true. The posterior probability in this case is equal to the prior probability, which is independent of the LR, and therefore the performance of the neutral method is always the same for different sets of LR values. 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 it is presented here (Equation 5), ECE is measured in units called *bits*, according to information theory [24].

In the ECE plot, the value of ECE (solid curve) where $\log_{10} \frac{Pr(\theta_p)}{Pr(\theta_d)} = 0$ (*i.e.*, where $Pr(\theta_p) = Pr(\theta_d) = 0.5$) can be viewed as a summarizing measure of performance, and it is called C_{llr} (see [20, 25] for details). The lower the value of C_{llr}, the better the overall performance of the set of LR values. The corresponding value of C_{llr} after PAV (*i.e*, the value of the dashed curve in the ECE plot where $\log_{10} \frac{Pr(\theta_p)}{Pr(\theta_d)} = 0$) is called C_{llr}^{min}. C_{llr} expresses the overall performance of a set of LR values as a single scalar number. C_{llr}^{min} measures the discriminating power of a set of LR values in a single scalar number, being $C_{llr}^{min} = 0$ for totally-separated LR values when θ_p and θ_d are respectively true. $C_{llr}^{cal} = C_{llr} - C_{llr}^{min}$ is a measure of the calibration loss of the experimental set of LR values.

As a summarizing example, the value of C_{llr} in Figure 2 for a value of (*i.e.* the value of ECE where $\log_{10} \frac{Pr(\theta_p)}{Pr(\theta_d)} = 0$) is 0.73 bits. This means that before the evidence was analyzed the loss of information (given by the neutral, dotted curve) was 1 if the prior probability is $Pr(\theta_p) = Pr(\theta_d) = 0.5$. After the evidence is analyzed, the information loss is reduced to 0.73. This values can also be viewed as the information needed to obtain certainty about the true values of θ_p and θ_d in the experimental set.

2.6. Software

Calculation of LR values were conducted by software 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[5].

[5] http://niko.brummer.googlepages.com/focal

3. Results and Discussion

This work is based on results of SEM-EDX analysis for a classification problem, where the hypotheses of interest are as follows:

- θ_p: the recovered glass fragments come from containers containers (category p).
- θ_d: the recovered glass fragments come from car and building windows (category cw).

Glass from car and building windows is typical in forensic cases. As highlighted before, glass samples from car windows (c) and building windows (w) categories need to be treated with care due to the very similar elemental content of these two categories. This is a consequence of the similarities in the way they are manufactured (a float glass manufacturing method). In this work we group them together in a single category. Also, glass from containers (p) are also common in forensic practice.

3.1. Descriptive statistics

Descriptive statistics of the observed elements in the considered glass databases are shown in this section. In Fig. 3 the marginal distribution of the different variables are given separately for containers (category p) and car and building windows (category cw). It is observed that the distribution of some of the variables is significantly different among categories, suggesting a potential usefulness in classifying categories from them. On the other hand, Table 1 shows the partial correlation among different variables, computed from the database. It is observed that some of the variables are highly correlated, as it could be expected given the strong dependencies of different elemental concentrations. This means that models should take into account this dependencies in order to properly model the behavior of the data in a probabilistic way. Thus, multivariate models such as the one proposed here are encouraged.

3.2. Experiments on variable selection

3.2.1. Variable selection

Calculation of a full model for glass classification, which takes into account all variables, requires the estimation of the probability density function under each of two propositions, θ_p and θ_d. For example in this article, glass fragments could be described by seven variables (results of SEM-EDX analysis), then it is necessary in a full model to reliably estimate 7 means, 7 variances, and 21 covariances, which is difficult from a small sample.

The approach in this article is selecting a small amount of variables for building the model, according to the ECE measure of performance. This known as feature selection, and represents a form of dimensionality reduction which allows a more reliable modelling of small datasets. In order to select the more convenient variables for the classification problem, we have performed an exhaustive feature selection strategy for all the possible combination of 1, 2 and 3 variables (namely univariate, bivariate and trivariate variable sets) from the available variables Al, Ca', Fe', K', Si', Mg' and Na'. As the number of possible variables is 7, there will be 7 possible univariate variable sets, $(7 \times 6)/2 = 21$ possible bivariate sets and $(7 \times 6 \times 5)/(3 \times 2) = 35$ possible trivariate sets. As a consequence, $7+21+35 = 63$ variable sets were tested. For each of the tests, the C_{llr} performance measure has been computed, and the sets have been ranked according to its value. The variable sets presenting a lower value of C_{llr} will be considered the best ones.

3.2.2. Experimental protocol

Due to the time-consuming process of obtaining elemental compositions of glass objects using the SEM-EDX method, the number of **glass-objects** in the available database (278) is rather limited. Thus, a jackknife procedure was employed to use the available data efficiently while obtaining reliable results. **With Jackknife, for each comparison of glass objects in order to obtain a LR value, the rest of objects is used for training the LR models. Thus, for** each iteration of the jackknife, one sample for the whole set of 278

glass objects was separated, creating a 1-object test set. The rest of 277 samples were used as the tuning set for the LR model. Thus, parameters of the numerator of the LR in Equation 3 were estimated using all the elements of the 277-object tuning set which belonged to the p category (because when θ_p is true the p category is assumed). On the other hand, parameters of the denominator of the LR in Equation 3 were estimated using all the elements of the 277-object tuning set which belonged to the cw category (because when θ_p is true the cw category is assumed). Finally, the elemental composition of the 1-object test set was used as the glass to be classified, which is represented as $\mathbf{y}$ in Equation 3, and we produce a LR value. The jackknife process is iterated to produce 278 LR values for the 278 1-object test sets possible. According to Section 1.4, in the case where $\mathbf{y}$ is obtained form an object belonging to p category (*i.e.*, θ_p is true) the LR should be above 1, and any LR value below 1 is considered to yield misleading evidence. Similarly, when $\mathbf{y}$ is obtained form an object belonging to cw category (*i.e.*, θ_d is true) the LR should be below 1, and any LR value above 1 is considered to yield misleading evidence.

3.2.3. Results

First, Figure 4 shows the (sorted) C_{llr} performance of the LR values obtained by the multivariate LR model (Section 2.3) is shown for all the 63 possible univariate, bivariate or trivariate combinations of variables. It is seen that many of the combinations of variables obtains values of $C_{llr} > 1$ (highlighted in Figure 4), which means that ECE$>$ 1 where $\log_{10} \frac{Pr(\theta_p)}{Pr(\theta_d)} = 0$. According to Section 2.5, this means that, with the selection of those variables, the model behaves worse than the neutral method (which obtains $C_{llr} = 1$ at $\log_{10} \frac{Pr(\theta_p)}{Pr(\theta_d)} = 0$). As a consequence, combinations of variables having $C_{llr} > 1$ behave worse on average than not evaluating the evidence at all (neutral method), and therefore those variables combinations should not be used for evidence evaluation. It is also seen in Figure 4 that the three best combination of variables in the sense of C_{llr} are (Si', K', Ca'), (Na', Si', Ca') and (K', Ca'). The ECE plots of these three cases is shown in Figures 5, where it is seen that the values of ECE (red,

solid curve) remain much lower than the ECE of the neutral method (black, dotted curve) for all the values of the prior log-odds in the x-axis. This means that the behavior of the model when such combination of variables are used is satisfactory for all possible values of the prior probabilities. This is an important issue, since the prior probabilities are not usually known by forensic scientists, and therefore the models used for evidence evaluation should be informative for all the possible prior probabilities. This cannot be seen with the use of C_{llr} alone, because it only focus on one prior probability, namely $\log_{10} \frac{Pr(\theta_p)}{Pr(\theta_d)} = 0$, which means $Pr(\theta_p) = Pr(\theta_d) = 0.5$. As a consequence of the analysis, we conclude that the best sets of variables to be used for glass classification in the context described here are (Si', K', Ca'), (Na', Si', Ca') and (K', Ca'), which are informative for all possible values of the prior probabilities. In order to give a clearer insight of what is measured with ECE, the histograms of the LR values for each of the three best sets of variables are also shown in Figure 5, just below their corresponding ECE plots. It is seen that the histograms presents a nice separation between LR values obtained from glass of p category (θ_p is true) and glass of *cw* category (θ_p is true). Moreover, the distributions are coherent in the sense of misleading evidence, because most of the former are greater than 1, and most of the latter are lower than 1. Therefore, the corresponding set of variables is suitable for glass classification, which is represented by a good value of ECE (low). Note that when histogram distributions when θ_p and θ_d are respectively true are completely separated, ECE plot after PAV (blue-dashed curve) in their corresponding ECE plots is zero, according to what is stated in Section 2.5. In order to further clarify what is being measured by ECE plots, we analyze three cases of combinations of variables where the value of C_{llr} is greater than 1, suggesting that those methods should not be used for evidence evaluation. Figure 6 shows three of these *bad* variable combinations, namely (Mg'), (Na', Ca') and (Na', Si', Fe'). In the case of (Mg') (Figure 6(a)), it is seen that the value of ECE is much higher than 1, especially for negative values of $\log_{10} \frac{Pr(\theta_p)}{Pr(\theta_d)}$. In the corresponding histogram, it is clearly seen that there are LR values lower than 0.001 when θ_p is true, which would give a high contribution to

ECE in Equation 5, and thus increasing their value. Moreover, these values will be more important for ECE when $Pr(\theta_d) >> Pr(\theta_p)$. This has the following interpretation: a strongly misleading LR value when θ_p is true yields to a very low value of the posterior odds in Equation 2. As the final decision will be usually taken from the posterior odds, even if the prior odds favor θ_p (*i.e.*, if they are higher than 1), the posterior odds may be lower than 1, and therefore they will lead to a bad decision. This is interpreted in terms of ECE as a loss of information about the correct hypothesis. Note also in Figure 6(a) that the Mg' variable presents a moderate discriminating power, since ECE after PAV (blue, dashed curve) is relatively far from the neutral ECE performance (dotted curve). This is also evidenced by the slight separation of the histograms when θ_p and θ_d are respectively true. However, due to the presence of strongly misleading evidence, this variable configuration is not informative. In the case of (Na', Ca') (Figure 6(b)) and (Na', Si', Fe') (Figure 6(c)) the problem is different: the discriminating power of these techniques is extremely poor, because the ECE after PAV (dashed curve) and the neutral ECE (dotted curve) in the ECE plots are almost the same for both variable sets. Thus, as ECE is always greater than ECE after PAV, the overall value of ECE (solid curve) is higher than 1, mainly due to a poor discrimination loss. Note that this discrimination loss is clearly seen in the histograms, where $\log_{10}$ of the LR values are totally mixed when θ_p and θ_d are respectively true. We analyze the relationship among discriminating power (separation of histograms) and calibration in Figure 7, where we show three variable combinations for which the discriminating power is good (ECE after PAV, dashed curve) but the the ECE value is high. All these cases are univariate, namely (Fe'), (K') and (Si'). From the observation of histograms, it is seen that in all the variable combinations the problem is the same: although the separation of LR values when the glass actually belongs to p (θ_p is true) and cw (θ_d is true) is fails nice (total separation for the (Fe') and (K') cases), the absolute value of the $\log_{10}(LR)$ is very close to 0, indicating LR values very close to 1 in almost all cases. It is said that this kind of LR values yield *weak evidence*. Thus, according to Equation 2, the posterior odds will be dominated

by the prior odds, and the evidence (glass analysis) is giving little information. This information loss is represented by a high value of ECE (solid curve), close to the neutral ECE (dotted curve). Thus, ECE clearly identifies when evidence is giving information useful to the decision process according to Equation 2. Note that a LR value close to 1 will contribute moderately to ECE in Equation 5, since the logarithmic scoring rule will have a moderate value for such a LR value (Figure 1 and Equation 4). **Moreover, iron and potassium are in general detected in samples from c, p and w categories at level close to SEM-EDX technique detection level. Therefore, an application of variables based on these elements (K' and Fe') should not be recommended for solving a classification problem.** The only way of reduce the contribution **of LR value close to** 1 to ECE is increasing the strength of the LR value while supporting the right hypothesis. A further interpretation is in order here: as the variable combinations which exhibit this behavior are univariate, we may explain the weakness of the evidence by the fact that just one variable gives some discriminating information, but it may be increased by using several variables. For instance, the best performing model in this work is trivariate, namely (Si', K', Ca') in Figure 5(a), including information both of variables (K') and (Si'), as well as (Ca'). As it can be seen in the histogram of Figure 5(a), the absolute value of the $log_{10}(LR)$ is significantly higher than 0, yielding stronger evidence than for univariate models. In order to compare ECE to other classical measures of performance, we show in Figure 8 the rates of misleading evidence for LR= 1 for all the 63 possible univariate, bivariate and trivariate combinations of variables. the rates of misleading evidence are defined as the proportion of LR values which are lower than 1 when θ_p is true, or LR values which are higher than 1 when θ_d is true, and is a popular measure of overall assessment of LR values [8, 26]. The plot is ordered according to C_{llr} value, in the same way as in Figure 4. By comparing Figure 8 and Figure 4, it is easily seen that when C_{llr} gets lower the rates of misleading evidence tend to decrease. Therefore, by optimizing ECE and its particular value C_{llr} the rates of misleading evidence tend to reduce. This result justify the use of ECE even when

rates of misleading evidence are used as main performance measures. Finally, we analyze which elements tend to give better results in term of ECE. Figure 9 shows the box plots of the C_{llr} and C_{llr}^{min}, grouped in different variables. Thus, each individual plot represents the distribution of C_{llr} (respectively C_{llr}^{min}) values for all the combination of variables which includes the variable in the x-axis. For instance, in Figure 9(a), for variable Al' in the x-axis, it is shown the distribution of the C_{llr} values of variable combinations including Al'. It is seen that variable combinations including K' and Na' variables tend to give significantly better C_{llr} and C_{llr}^{min} values. On the other hand, variable combinations including Al' variable clearly tend to present bad results. Thus, for classification of glass traces, it will be worth considering not taking into account the Al' variable. This is in accordance with Figure 4, where all the variable combinations including the Al' variable perform bad.

4. Conclusions

In this article, a selection technique of the most convenient features for glass classification using the SEM-EDX technique has been proposed in the context of forensic evaluation of the evidence. The classification is done by the use of a likelihood ratio model in a Bayesian inferential context. The model has been previously proposed in [14]. The aim is reducing the complexity of the dataset in order to make the models more robust in sparse data conditions. This is a typical situation in forensic glass analysis nowadays as glass analysis is in general time consuming and large databases are not currently publicly available. In order to perform the selection of the best variables, a measure of goodness has been proposed, previously used in forensic glass comparison problems [8], namely Empirical Cross-Entropy presented in the form of so-called ECE plots. They allow to easily represent not only the discriminating power of the model in use but also its calibration, which indicates whether the likelihood ratios given by the model are interpretable in a probabilistic way.

Results show that a trivariate model based on the ratios of elemental con-

centration of silicon, potassium and calcium with respect to oxygen (namely (Si', K', Ca') variable set) presents the best performance, obtaining almost total separation of glass objects coming from containers (p category) and car or building windows (cw category). This model also presents good calibration, and therefore it gives likelihood ratios which are interpretable in a probabilistic way in a forensic Bayesian inferential process. This analysis shows the adequacy of the ECE measure as a powerful evaluation tool to determine the goodness of the models used in classification tasks. Given the usefulness of ECE plots, a software for ECE analysis is publicly available by one of the authors.

Acknowledgments

Dr. Daniel Ramos's research is co-funded by the Universidad Autonoma de Madrid and the Comunidad Autonoma de Madrid under project CCG10-UAM/TIC-5792.

Dr. Grzegorz Zadora research is co-funded by the Institute of Forensic Research under project IX/K/2009-2010.

References

1. G. Zadora, Z. Brozek-Mucha, SEM-EDX - a useful tool for forensic examinations, Material Chemistry and Physics 81 (2003) 345–348.

2. 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.

3. C. Latkoczy, S. Becker, M. Ducking, D. Gunther, J. Hoogewerff, J. Almirall, J. Buscaglia, A. Dobney, R. Koons, S. Montero, G. van der Peijl, W. Stoecklein, T. Trejos, J. Watling, V. Zdanowicz, Development and evaluation of a standard method for the quantitative determination of elements in float

glass samples by la-icp-ms, Journal of Forensic Sciences 50 (6) (2005) 1327–1341.

4. 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.

5. 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).

6. G. Zadora, Classification of glass fragments based on elemental composition and refractive index, Journal of Forensic Sciences 54 (2009) 49–59.

7. G. Zadora, T. Neocleous, C. Aitken, Two-level model for evidence evaluation in the presence of zeros, Journal of Forensic SciencesTo appear.

8. G. Zadora, D. Ramos, Evaluation of glass samples for forensic purposes – an application of likelihood ratio model and information-theoretical approach, Chemometrics and Intelligent Laboratory Systems 102 (2010) 62–63.

9. C. G. G. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, John Wiley & Sons, Chichester, 2004.

10. I. W. Evett, Towards a uniform framework for reporting opinions in forensic science casework, Science and Justice 38 (3) (1998) 198–202.

11. A. W. F. Edwards, Likelihood, The Johns Hopkins University Press, Baltimore, MD, USA and London, UK, 1992.

12. 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.

13. 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.

14. G. Zadora, T. Neocleous, Likelihood ratio model for classification of forensic evidences, Analitica Chimica Acta 642 (1-2) (2009) 266–278. doi:10.1016/j.aca.2008.12.013.

15. J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, J. Ortega-Garcia, Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition, IEEE Transactions on Audio, Speech and Language Processing 15 (7) (2007) 2072–2084.

16. I. Evett, Expressing evaluative opinions: A position statement, Science and Justice 51 (2011) 1–2, several signatories.

17. R. Royall, On the probability of observing statistical misleading evidence, Journal of the American Statistical Association 95 (451) (2000) 760–768.

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. M. H. deGroot, S. E. Fienberg, The comparison and evaluation of forecasters, The Statistician 32 (1982) 12–22.

20. N. Brümmer, J. du Preez, Application independent evaluation of speaker detection, Computer Speech and Language 20 (2-3) (2006) 230–275.

21. 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.

22. 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, pp. 411–416.

23. B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, 1986.

24. T. M. Cover, J. A. Thomas, Elements of Information Theory, 2nd ed., Wiley Interscience, 2006.

25. N. Brummer, Measuring, refining and calibrating speaker and language information extracted from speech, Ph.D. thesis, School of Electrical Engineering, University of Stellenbosch, Stellenbosch, South Africa, available at http://sites.google.com/site/nikobrummer/ (2010).

26. D. Dessimoz, C. Champod, Linkages between biometrics and forensic science, in: A. K. Jain, P. Flynn, A. A. Ross (Eds.), Handbook of Biometrics, Springer, 2007.

5. Table

Table 1: Partial correlation matrix for the seven variables based on the variance-covariance matrix.

	Na'	Mg'	Al'	Si'	K'	Ca'	Fe'
Na'	1.000	0.181	-0.174	-0.691	0.304	0.367	-0.097
Mg'		1.000	0.104	-0.369	0.011	0.361	-0.096
Al'			1.000	0.146	-0.439	-0.024	0.224
Si'				1.000	-0.187	-0.786	0.032
K'					1.000	0.067	-0.068
Ca'						1.000	-0.040
Fe'							1.000

6. Figures

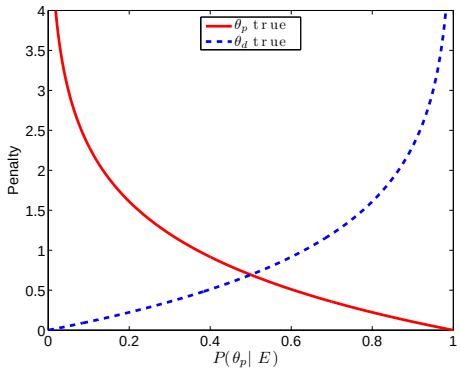


Figure 1: Logarithmic scoring rule. The x-axis represents the posterior probability of θ_p, which may be viewed as the "forecast" of about whether θ_p is true considering all the available knowledge in a given forensic case, including the evidence E.

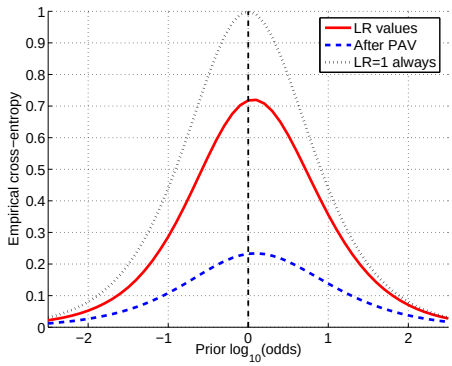
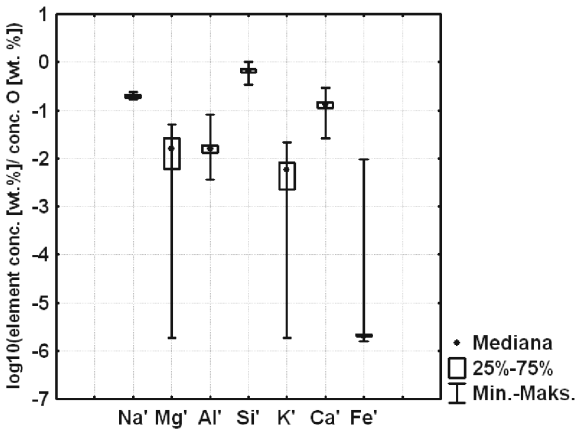
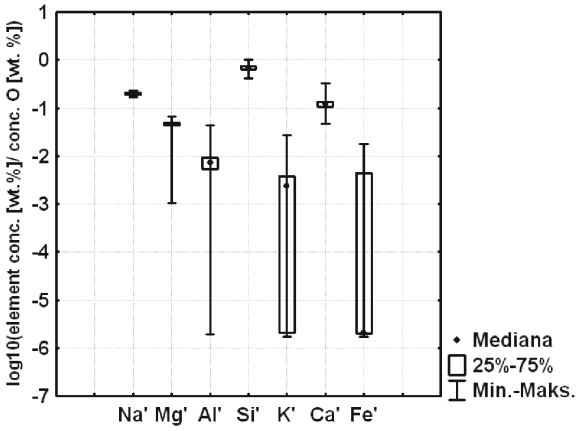


Figure 2: Example of ECE plot. C_{llr} is highlighted as the ECE (red curve) where $\log_{10} \frac{Pr(\theta_p)}{Pr(\theta_d)} = 0$ in the x axis (vertical, dashed line).



(a)



(b)

Figure 3: Box-plots showing marginal distributions of variables in each of the two glass categories (p - container, cw - car and building windows).

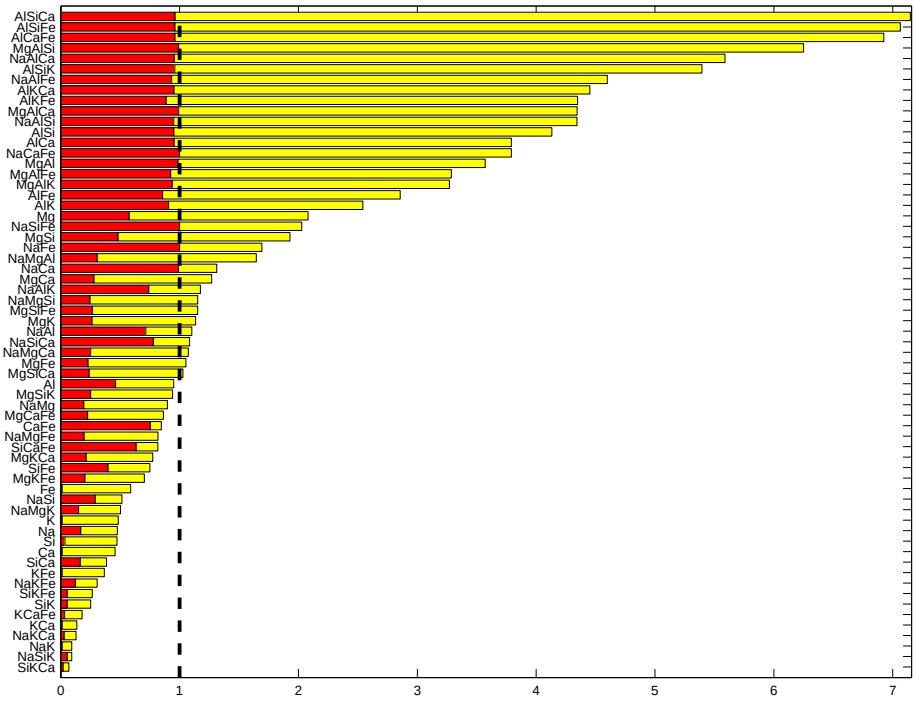
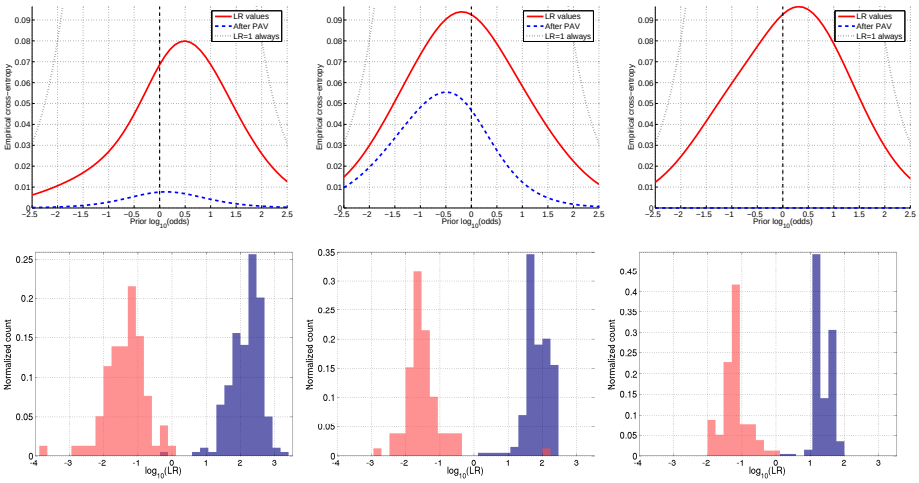
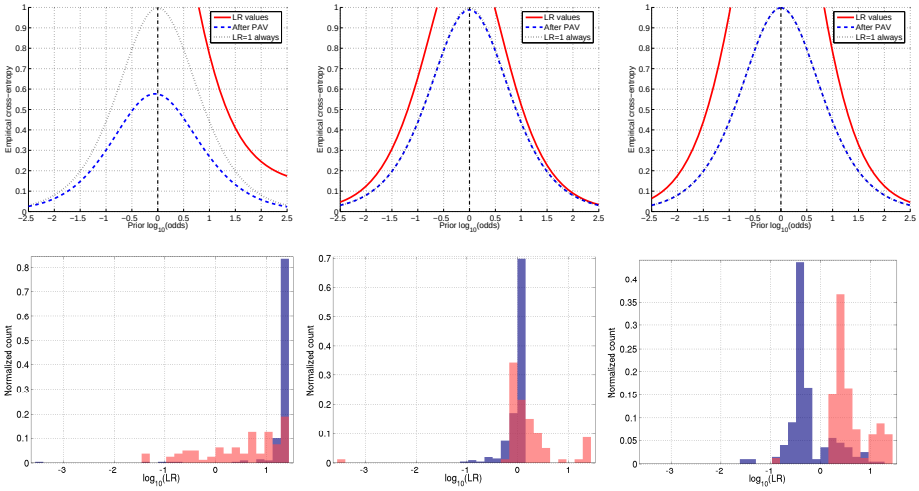


Figure 4: Performance, in terms of C_{llr} of the LR values obtained with the proposed multivariate model and all the 63 possible univariate, bivariate and trivariate combinations of variables. C_{llr} is the total bar for each combination of variables (darker red bar plus lighter yellow bar, both stacked together). It is decomposed as $C_{llr} = C_{llr}^{min} + C_{llr}^{cal}$, where C_{llr}^{min} is represented in darker red and C_{llr}^{cal} is represented in lighter yellow.



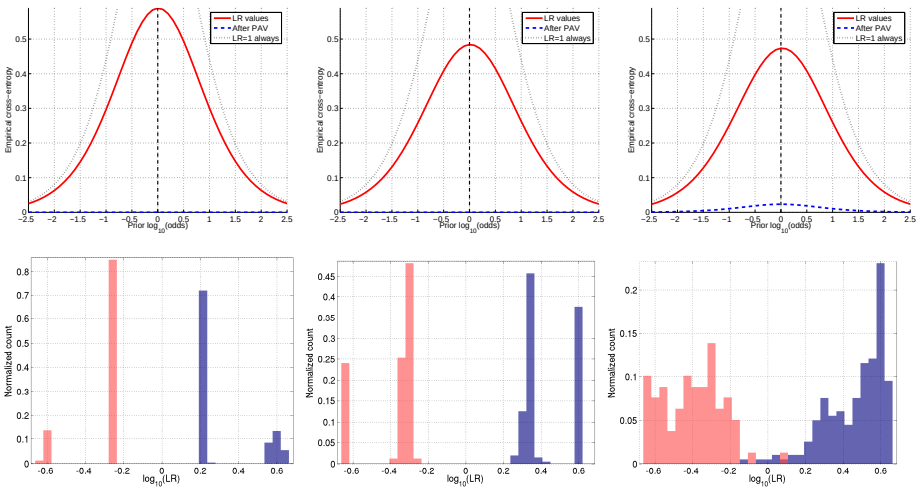
(a) (b) (c)

Figure 5: ECE plots (top) and histograms of $\log_{10}(LR)$ values (bottom) for the three best combinations of variables, *i.e.* the three combinations of variables presenting the best C_{llr} value: (Si', K', Ca') (a), (Na', Si', K') (b) and (Na', K') (c). Histograms are darker blue when θ_p is true (glass to classify actually belongs to category p) and lighter red when θ_d is true (glass to classify actually belongs to category cw).



(a) (b) (c)

Figure 6: ECE plots (top) and histograms of $\log_{10}(LR)$ values (bottom) for three combinations of variables presenting a bad, $C_{llr} > 1$ value: (Mg') (a), (Na', Ca') (b) and (Na', Si', Fe') (c). Histograms are darker blue when θ_p is true (glass to classify actually belongs to category p) and lighter red when θ_d is true (glass to classify actually belongs to category cw).



(a) (b) (c)

Figure 7: ECE plots (top) and histograms of $\log_{10}(LR)$ values (bottom) for three combinations of variables presenting total discrimination (ECE after PAV is 0), but moderately high C_{llr} value: (Fe') (a), (k') (b) and (Si') (c). Histograms are darker blue when θ_p is true (glass to classify actually belongs to category p) and lighter red when θ_d is true (glass to classify actually belongs to category cw).

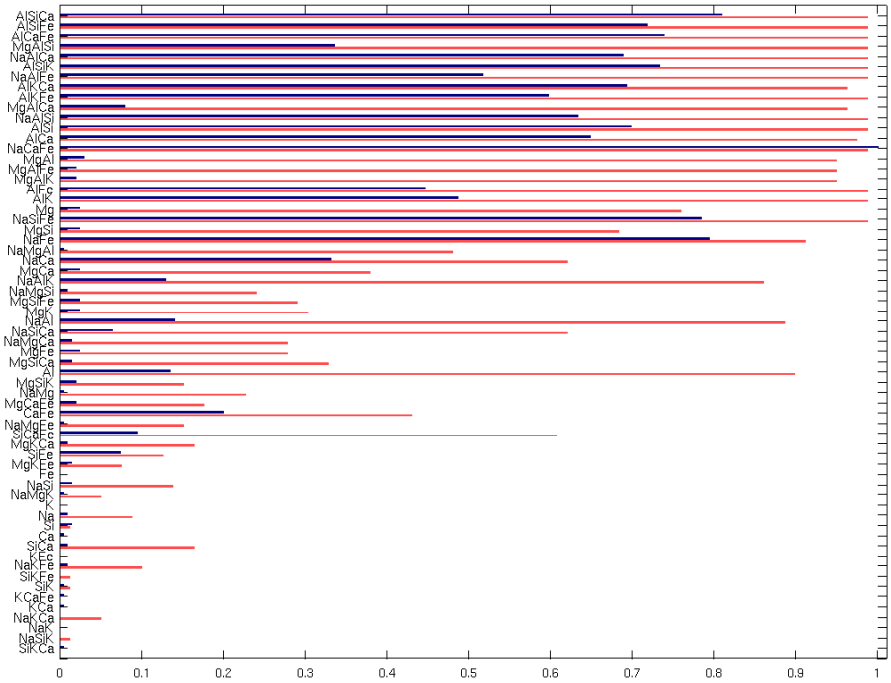
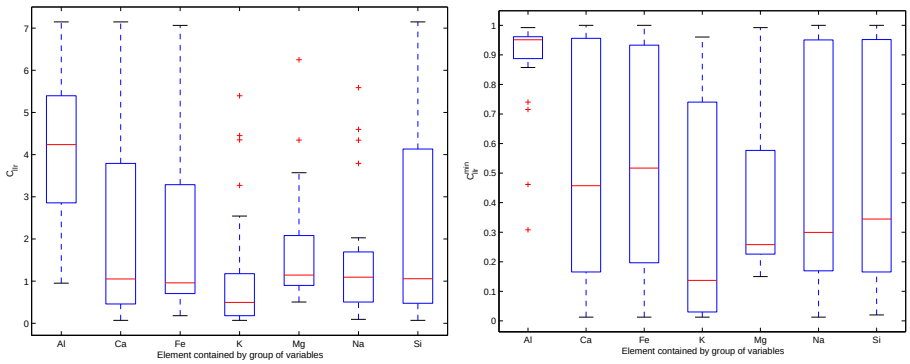


Figure 8: Rates of misleading evidence at LR= 1 for all the LR values obtained with the proposed multivariate model and all the 63 possible univariate, bivariate and trivariate combinations of variables. For each combination of variables, darker blue bar is the rate of misleading evidence for θ_p is true (proportion of LR values lower than 1 when glass to classify actually belongs to category p) and lighter red when θ_d is true (proportion of LR values higher than 1 when glass to classify actually belongs to category cw)..



(a) (b)

Figure 9: Box plots of C_{llr} (a) and C_{llr}^{min} (b) grouped by different elemental variables. For instance, variable Al' in the x-axis determines the box plot of the C_{llr} (a) or C_{llr}^{min} (b) values of variable combinations including Al'.