Local Ion Signatures (LIS) for the examination of comprehensive two-dimensional gas chromatography applied to fire debris analysis


DOI
10.1016/j.forc.2016.10.003

Publication date
2017

Document Version
Final published version

Published in
Forensic Chemistry

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Link to publication

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Local Ion Signatures (LIS) for the examination of comprehensive two-dimensional gas chromatography applied to fire debris analysis

Martin Lopatka a,f, Andjoe A. Sampat b, Steffan Jonkers b, Lawrence A. Adutwum c, Hans G.J. Mol d, Guido van der Weg d, James J. Harynuk c, Peter J. Schoenmakers b, Arian van Asten b,e,f,⇑, Marjan J. Sjerps a,f,e, Gabriel Vivó-Truyols b

Korteweg de Vries Institute for Mathematics, University of Amsterdam, Postbus 94248, 1090 CE Amsterdam, The Netherlands
Van ’t Hoff Institute for Molecular Sciences, University of Amsterdam, Postbus 94157, 1090 GD Amsterdam, The Netherlands
Department of Chemistry, University of Alberta, Cuming/Lemieux Chemistry Centre, 11227 Saskatchewan Drive, T6G 2G2 Edmonton, Alberta, Canada
RIKILT – Wageningen UR, Postbus 230, 6700 AE Wageningen, The Netherlands
Co van Ledden Hulsebosch Center (CLHC), Amsterdam Center for Forensic Science and Medicine, Postbus 94157, 1090 GD Amsterdam, The Netherlands
Netherlands Forensic Institute, Postbus 24044, 2490 AA Den Haag, The Netherlands

1. Introduction

1.1. Forensic relevance

Records kept by the Dutch Statistics Bureau (CBS) [1] indicate that 19.2% of the indoor fires responded to in 2013 were suspected (by fire department first responders) to be arson cases. Due to the immense societal cost inflicted by accidental and deliberate fires, substantial resources are invested in the forensic analysis of fire debris; a task requiring both forensic expertise and the capacity to perform sophisticated chemical analyses. Forensic fire investigation may involve the detection of ignitable liquid residues (ILR) in fire debris from the scene. Depending on the circumstances, such residues may constitute important evidence that arson was committed and can provide valuable tactical information.

Despite the maturity of Fire Debris Analysis (FDA) as a domain of forensic practice, detection and classification of ILR remains a very challenging task. This detection and classification process is complex in nature and requires extensive experience from the forensic chemist to carry out interpretation [2]. Several factors...
Complicate the analysis of fire debris, such as the complex composition of most ignitable liquids [3], the diversity of ignitable liquid products easily accessible to arsonists [4,5], the partial consumption and evaporation of ignitable liquids during the fire [6,7], the formation of interfering pyrolysis products from the substrate [8,9], the invasive processes employed by fire fighters to stop and extinguish the fire [10] and the physical, chemical and biochemical processes that can occur after the fire [11,12].

It has been shown that these complicating factors can lead forensic experts to erroneous conclusions for challenging fire debris samples [13]. The subjective nature of the data interpretation has serious implications for accreditation and harmonization efforts as well as for the experimental design of studies use to evaluate the efficacy of automated methods. In the study presented herein, the authors choose to equate the pre-burn addition of ignitable liquids (IL) to a sample with the ground truth presence of ignitable liquid residue (ILR) in the fire debris. The goal of this manuscript is to demonstrate preliminary steps towards a methodology for objective and automated detection and classification of ILR in fire debris samples using comprehensive two-dimensional gas chromatography. Continued development of such methods could assist forensic experts with data interpretation and improve the objectivity of fire debris analysis.

1.2. Background

Previous work has demonstrated that the Total Ion Spectrum (TIS) [14,15] calculated over an entire chromatographic separation may yield a very valuable feature space in which the detection and classification of fire debris can be successfully performed [16–18]. The TIS is computed by summing the intensity of each nominal mass-to-charge value over all measurements along the retention time axis of a chromatographic separation. This approach provides many desirable characteristics for the comparison of chromatographic signals. One advantage is the robustness of this method to retention time shifts, frequently observed in chromatographic data [15]. Another advantage is the inherent reduction in dimensionality that comes from aggregating the mass intensities.

Within the domain of FDA, many state of the art techniques are score-based and feature-based approaches [34,39,7]. Each characterizes the evidence in a distinctly different manner, making different modeling assumptions in addressing source identity problems. Score-based methods report on the rarity of observing a particular score between a pair of samples, thus reformulating the evidence in terms of the rarities of observing similarity metric comparing samples rather than the original features [34]. Feature-based LRs report on the rarity of observing particular features under competing hypotheses pertaining to the source of the sample [35]. Recent research has drawn attention to the fundamental differences in the statistical models that can be defined within the LR framework [39,40,41] and the practical differences observed with real chemical measurements [34,42].

The use of an LR framework in the assessment of samples requires the formulation of competing hypotheses specific to the level of forensic interest being examined. For detection of ILR (for an arbitrary sample, x) the hypotheses may be formulated as follows:

- \( H_0: \) sample x does not contain ignitable liquid residue
- \( H_1: \) sample x contains ignitable liquid residue

Conventionally the subscripts \( p \) and \( d \) are adopted to indicate the hypotheses of the prosecution and defence respectively. For the classification of ILR, the hypotheses are reformulated to pertain to the classification of a sample (x) into an arbitrary class of ignitable liquids (y) with at least 2 mutually exclusive options here denoted \( i \) and \( j \).

- \( H_0: \) sample x contains ILR belonging to the class \( y_i \)
- \( H_1: \) sample x contains ILR belonging to the class \( y_j \)

The use of an LR framework for assessing fire debris samples has the distinct advantage that it can not only be used to perform hard classifications (if required) but also to explicitly indicate the strength of the evidence on a quantitative scale. This is a very important advantage for forensic casework, where experts should refrain from making absolute claims pertaining to class membership.

The forensic context explored herein is the detection and subsequent classification of ignitable liquid residues into three classes of interest: White Spirits (WS), Gasolines (GAS), and Lamp Oils (LO). GC × GC–MS is used to improve the performance (compared to similar approaches operating on 1-D data) by providing chromatographically separated mass spectra as input and performing feature selection sensitive to the elution times of compounds of interest. This considerably extends the Total Ion Spectra (TIS)
approaches explored in previously published research for one-dimensional chromatography [15] which is used in feature reduction of our analytical measurements.

Despite a great deal of data reduction achieved by summation across the chromatographic space (as in the TIS approach), the dimensionality relative to the number of available samples complicates statistical modeling of the original features. Research on chemical profiling for source identity has demonstrated a relationship between the evidential value computed using feature-based LRs and the evidential value computed using source-independent score-based LRs [34]. The score-based approach is implicitly different than a feature-based representation of the LR wherein the rarity of observing the features themselves is calculated under the competing hypotheses [39,43]. Nonetheless, score-based LRs may act as a proxy for the same-source versus different-source relationship by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34]. It is due to this relationship that the score-based method is suitable as a feature selection method. By modeling the LR attainable by indirectly modeling the relationship between samples via a pairwise similarity metric and can indicate similar forensic assertions [34].

2. Materials and methods

2.1. Laboratory fire experiments

In order to produce a realistic data set in which the presence or absence of ILR is known as well as the type of ILR that was used as an accelerant, small-scale fire debris samples consisting of various substrates were created in a controlled manner. Experiments were conducted in a fire- and explosion-proof laboratory, built at the Netherlands Forensic Institute (NFI) for hazardous experiments. An experimental setup was designed that included a total of 155 measurements (provided as Supplementary Material). A total of 19 different substrates, consisting of furniture and building materials commonly found in homes, were used to conduct the experiments. Out of the 155 measurements, 110 consisted of fire debris and 45 were analyses of neat ignitable liquids. More information regarding the brand and corresponding ASTM categories of the specific ignitable liquids used can be found in the Supplementary Material.

For 20 experiments only a single type of substrate was burned, 45 experiments included a combination of substrate materials without the addition of ignitable liquid, and 45 experiments were performed with exactly the same substrate composition and the addition of ignitable liquid. For each burn condition, a total of 300 grams of material was used. In each of the experiments (excluding single substrate burns) 60% of the total mass of material represented the primary substrate; the remaining 40% was comprised of 4 types of secondary substrates. Three types of ignitable liquids were used. For each type of ignitable liquid 15 different bottles, representing different brands, were purchased form different retail locations throughout the Netherlands.

Burn experiments were performed on single-use, uncoated, stainless-steel plates (345 × 240 mm). For experiments involving ignitable liquid (IL), 15 mL of IL was applied in three parallel lines using a glass Pasteur pipette. After allowing the IL to absorb for five minutes, the samples were ignited using a gas torch. The fires were left to burn freely for 8 min and extinguished with water using a spray bottle. After allowing the debris to cool (approximately 30 min), the entire sample was manually collected in glass jars.

2.2. Analytical methodology

Fire debris samples were first heated for 4 h at 70 °C before sampling 0.5 mL of headspace. This headspace sample was analysed according to a GC-FID screening procedure to attain a preliminary survey of volatiles present in the sample’s headspace. The maximum signal intensity observed in the screening analysis was used to determine an appropriate headspace volume for the extracts used in subsequent GC × GC–MS analysis. These were prepared by drawing headspace through a sorbent tube containing activated charcoal using a 100 mL syringe.

GC × GC–MS analyses were carried out on an Agilent Technologies 6890 N GC System with a LECO dual-stage, quad-jet thermal modulator coupled to a Pegasus III ToF-MS (LECO, St. Joseph, MI, USA). An Agilent DB-1 (100% dimethyl polysiloxane) first-dimension column (30 m × 250 μm i.d.; 0.5 μm film thickness) was coupled to an Agilent DB-17 ([50%-Phenyl]-methyl polysiloxane) second-dimension column (1 m × 100 μm i.d.; 0.2 μm film thickness). A modulation time of 4 s was used and detection was performed covering a 35–500 m/z range.

Full details of the analytical methodology, separation program, and suppliers of all chemical reagents can be found in the Supplementary Material.

2.3. Feature reduction

Initial data processing steps are required to reduce the dimensionality of a single measurement signal while maintaining information that may be valuable for the forensic investigation. The reduced form representation (RFR) entails a reduction in the number of features (per measurement) from [800 × 1185 × 466] variables to a consolidated data space of [t × 466] variables where t Local Ion Signatures (LIS) are computed from rectangular tiles dividing the entire chromatographic space into t equally sized tiles as depicted in Fig. 1. This approach, inspired by earlier work using the TIS [15–18], strives to include some valuable retention time information while still benefitting from the robust and informative reduced representation of the sample in this manner.

A LIS may be calculated for an arbitrary region in the chromatographic space by summing the intensity values per a particular nominal m/z over all of the points within the given region resulting in (for the data used here) a vector of [1 × 466] values for each of t regions specified. The local vectors are then concatenated in an order corresponding to the retention time yielding a single vector of [1 × (t × 466)]. The number of tiles t may easily be extended to an arbitrary number of regions, keeping in mind that smaller tile regions increase the risk of misalignment between measurements between chromatographic runs. To demonstrate the utility of this strategy we divide the chromatographic space into only 12 large tiles (based on retention time) meaning that the dataset resulting from all burn experiments discussed in this manuscript can be represented as a [155 × 5592] matrix. As seen in Fig. 1, we may capture compounds eluting at drastically different relative retention times per dimension with a simple grid of [4 × 3] tiles. It is important to note that the selection of a tiling strategy is of paramount importance and should be considered in the context of the specific samples and include domain knowledge of experts.

Due to the additive effect of summing over such a large number of data points, a substantial contribution is observed in the signal due to the accumulation of noise in the LIS, meaning that values of 0 are not possible in the signal even when no analytes are present. To remove some of the effects of the noise, each LIS vector is normalized by subtracting the minimum intensity observed per ion from the LIS as per Eq. 2, and subsequently divided element-wise with the adjusted TIS value as per Eq. (1). Thus for the tile t and the ion i:

\[
\text{lis}_{t,i} = \frac{\text{lis}_{t,i} - \text{lis}_{t}}{\text{tIS}}
\]
The RFR of a chromatogram may thus be described as the signal. The TIS is normalized in exactly the same manner, described.

Each normalized LIS ($lis_i$) corresponds to a single vector of 466 quantities calculated per tile for each of $t$ tiles that a chromatogram is divided into. The RFR for each chromatogram $i$ results from concatenating the $lis_i$ values such that: $RFR_i = [lis_{i1}, \ldots, lis_{it}]$. A practical limitation of the RFR is the very large discrepancy between low abundances of trace compounds, and the abundances of very common ions that may originate as fragmentation products of larger molecules. To reduce the effect of ions that are universally common in all tiles, each LIS vector is subsequently divided by the normalized TIS, see Eq. (1). The TIS is calculated as per the procedure described in [15] but summing along both retention time dimensions. For GC × GC this equates to summing the modulated signal. The TIS is normalized in exactly the same manner, described in Eq. (3). The RFR of a chromatogram may thus be described as the ratio of the local ion quantities relative to the total quantities of those ions in the whole chromatogram. This method of normalization has the desirable characteristic that a new sample may be independently normalized for comparison against a reference collection without depending on parameters determined from the reference collection.

2.4. Feature selection

Further dimension reduction is required to perform classification and evidential interpretation tasks on this data set. A score-based LR approach can be applied as a powerful tool for the selection of features of interest. A subset of the total data reserved for parameter tuning partition is used to independently evaluate the efficacy of each feature in distinguishing between same-source and different-source comparisons. When the detection of ignitable liquid residue is being investigated, then same-source pairwise comparisons refer to all comparisons made between samples containing any ignitable liquid residue of any type as well as all comparisons made between samples not containing ignitable liquid residue. Different-source comparisons refer to all comparison made between ILR present and ILR absent samples. We may therefore formulate a univariate score-based LR model for the general assessment of same-label versus different-label pairwise comparisons between two arbitrary samples $i$ and $j$ where $i \neq j$. Based on those classifications such a system would be described by Eq. (4). Here, the evidence is simply the absolute difference between samples, i.e. $|RFR_{ij} - RFR_{ij}|$ measured per feature $f$ where, $f = 1, \ldots, 5592$. The density of the absolute differences observed under each of the two competing hypotheses (for our data) may be described by a univariate normal probability density function $g(.)$.

$$LR_{ij} = \frac{g(|RFR_{ij} - RFR_{ij}|)}{g(|RFR_{ij} - RFR_{ij}|)}$$

where $label$ may mean ILR present/absent or a particular class of ILR, depending on the forensic level of interest. For each variable $f$ in the RFR of the chromatogram we perform the procedure described above. Training samples are compared based on their absolute difference; this process is performed exhaustively for all pairwise comparisons possible. For the forensic classification task the sense of intra-class similarity corresponds to pairwise comparisons performed between samples labels in terms of their ILR content. Adopting a labeling scheme based on the specific types of ignitable liquids: WS, GAS, and LO, any comparisons between a GAS sample and another GAS sample contributes to the distribution of same-label comparisons. Likewise, comparisons...
between different class ignitable liquids (including comparisons with ILR absent samples) count towards the different-label comparisons when generating the score-based LR model.

This entire procedure is repeated once per feature in the LIS representation, leading to pairs of univariate distributions describing the observed absolute differences between samples of the same and of different type, for only one feature at a time. This means that 5592 score-based LR systems have been constructed for the detection of ILR and another 5592 for the forensic task of ILR classification. Having also set aside a data partition for the tuning of parameters, the performance of these LR systems in distinguishing between similar label and dissimilar label measurements with a known ground truth is evaluated. Using the known ground truth and the calculated LRs it is possible to evaluate the efficacy of each variable (f) in distinguishing between types. The metric by which each feature is examined is the log likelihood ratio cost or CLLR [41], calculated by Eq. (5):

$$C_{lr} = \frac{1}{2} \left( \frac{1}{N_{SL}} \sum_{i=1}^{N_{SL}} \log_2 \left( 1 + \frac{1}{LR_{SLi}} \right) + \frac{1}{N_{DL}} \sum_{i=1}^{N_{DL}} \log_2 \left( 1 + LR_{DLi} \right) \right)$$  (5)

where $N_{SL}$ is the number of comparisons made between samples with the same label and $N_{DL}$ is the number of different-label comparisons performed. $LR_{SL}$ and $LR_{DL}$ are the corresponding likelihood ratios for same-label and different-label comparisons respectively. The CLLR metric imposes a penalty based on the magnitude of LRs inconsistent with the known ground truth [41]. Deviation from a CLLR of 0 is indicative of poor performance of a system for reporting evidence. The CLLR accounts for not only the incidence of misleading evidence but also the magnitude of the erroneous LRs reported. In this case, each system of LRs being evaluated comprises of a single feature in the RFR feature space.

Thus the evaluation of potential utility for each feature is based on the CLLR achieved. Fig. 2 shows the CLLRs for an arbitrary Monte Carlo iteration. The 5592 features in the RFR are listed along the x-axis. For the stratified CV conditions used here 10 data partitions are defined, therefore the parameter tuning partition consists of 15 samples and their corresponding 105 pairwise comparisons are used in computing a single CLLR per feature. The features are now ranked based on their corresponding CLLR value and the top ranked features (i.e. those with the smallest CLLR values) are retained to use in subsequent modeling steps. Any subset of the features selected may be directly related to both a spatial position in the original two-dimensional chromatogram (specific to a tile) and a specific nominal m/z value. This is a valuable characteristic when interpreting the evidential value of any assertions made on the basis of these features and maintains a clear and intuitive relationship with the chemical information contained in the original data (See discussion in Section 4.1). The application of a CLLR measurement on each individual feature does not address the multicollinearity problem frequently encountered when performing statistical analysis on chemical trace data [44]. Nonetheless, the generalization into a series of univariate score-based LR models allows a pragmatic basis for further feature selection.

2.5. Ignitable liquid detection and classification

A classification subspace is generated by discarding features not ranked in the top 25 positions by the procedure described in Section 2.4. The number of features used must strike a fine balance in order to retain the maximum amount of information possible from the original data while ensuring that sufficient samples are available to reliably calculate the covariance parameter needed in the subsequent processing steps. Therefore, the choice of 25 features in our case is dictated by the number of samples available. The top 25 features of the training partition data are then projected onto a subspace defined by performing LDA [44]. If the forensic level of interest pertains to ignitable liquid detection, the competing hypotheses may be represented as two classes. Ground truth labeling of each measurement is performed according to the presence or absence of ignitable liquid residue.

$$y_i = 1$$ if no ignitable liquid was used in the experimental burn.

$$y_i = 2$$ if any type of ignitable liquid was used as an accelerant in the experimental burn.

LDA approaches a two-class classification problem by approximating a conditional probability density function for each class under the assumption that the two classes have identical covariance parameters and distinct means and that both are normally distributed [44].

The classification of a new measurement ($RFR_{new}$) is a function of its position in the projected space relative to some decision boundary. Under the presupposition of equal variance the class membership of a new point is simply defined as the minimum distance to a class mean in the projected space. The LDA projection will be a linear combination of

![Fig. 2. CLLR value calculated per feature in reduced LIS representation of sample #93.](image-url)
the features selected by their CLLR ranking and allows for a further reduction in dimensionality while minimizing the information loss incurred. Furthermore, the multicollinearity among these features is accounted for by their relative contribution to the projected space. The projected training samples are then modeled by a Gaussian probability density functions describing the different classes from the projected data without the added constraint of homoscedasticity [45].

An example depicting the projection of new data points to be classified, in terms of the two-class ignitable liquid detection task, can be seen in Fig. 3. Since the reduced dimensional hyperplane is defined in terms of a linear combination of the original features, the ratio of the densities between the two classes may also be expressed in terms of a feature-based LR, wherein the evidential term is now the representation of a new measurement in the one-dimensional feature space defined by the LDA projection for any given data point. The LR for a given classification can then be expressed by Eq. (6).

\[
LR = \frac{g \left( \text{projection}(RFR_i) | \mu_{y=1}, \sigma_{y=1} \right)}{g \left( \text{projection}(RFR_i) | \mu_{y=2}, \sigma_{y=2} \right)}
\]  

(6)

The competing hypotheses are now represented by the parameters defining a Gaussian distribution representing that class of data points on the LDA projected line, where \( g(.) \) is the Gaussian probability density function.

The extension into multi-class LDA is described in detail [44] and is also described alongside a computational implementation that formed the basis of the one used in the calculation of these results [46]. Essentially, the extension can be seen as classifying a number of samples into three distinct categories based on selecting a two-dimensional projection of the original data maximizing the variance between classes while minimizing the variance within classes [44]. The LR formulation, in analogy with the two-class formulation described in Eq. (6) may be reformulated as the likelihood of a given point belonging to a particular class versus the likelihood of it belonging to another specific class, as seen in Eq. (7) for class 1 versus class 2.

\[
LR = \frac{g \left( \text{projection}(RFR_i) | \mu_{y=c_1}, \Sigma_{y=c_1} \right)}{g \left( \text{projection}(RFR_i) | \mu_{y=c_2}, \Sigma_{y=c_2} \right)}
\]  

(7)

In the case of a three class formulation the hypotheses are formulated as follows. This formulation explicitly disregards the possibility of additional classes and false positives, however, may be extended to any number of classes, so long as they are explicitly defined and mutually exclusive. For any new classification in a 3 class problem, 2 LRs are produced for the hypotheses specified in Table 1. If one were to model the denominator as a conjunction of \( H_a \) and \( H_b \) an implicit prior probability is unavoidably introduced pertaining to the relative rarity of each IL class in the relevant population. Results reported here focus on the likelihoods when evaluating a one-versus-one formulation of the competing hypotheses.

In the LDA subspace, the likelihood of observing the evidence (projected features values) under each of the \( H_a \) cases must be computed. An example, representing new samples projected into a two-dimensional hyperplane for the three-class ignitable liquid residue classification problem is depicted in Fig. 4.

### Table 1

<table>
<thead>
<tr>
<th>Hypotheses for the classification of ignitable liquid residue.</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS ( H_a ): sample ( x ) contains ILR of type WS</td>
</tr>
<tr>
<td>GAS ( H_b ): sample ( x ) contains ILR of type GAS</td>
</tr>
<tr>
<td>LO ( H_c ): sample ( x ) contains ILR of type LO</td>
</tr>
</tbody>
</table>

3. Results

In order to guard against overfitting, all results reported are based on aggregate findings in a series of repeated Monte Carlo iterations using a stratified cross-validation (CV) design with 10 folds. The data is partitioned such that every Monte Carlo iteration contains the full 10 step cross-validation run where: 1/10 of the data (proportionally sampled per experiment type and ignitable liquid type) is assigned to be a testing partition, 1/10 is assigned to be a parameter tuning partition and the remaining 8/10 are used as a training partition. For one Monte Carlo iteration to complete, each of the 10 partitions is used exactly once as the testing partition. An algorithmic representation of the Monte Carlo workflow is presented in the Supplementary Material.

Example chromatograms from each of the three IL classes focused on within the context of this research are presented in Fig. 5 demonstrating the complexity of the samples and abundance of potential chemical information that acts as a starting point for both manual examination by forensic chemists and automated, statistically driven methods. By using GC × GC to drastically
increase the number of resolvable peaks in the resulting chromatograms an abundance of data is created to facilitate the selection of informative features for the detection and classification tasks explored.

3.1. Detection of ignitable liquid residue

In the first step, the detection of ILR is performed. Samples from the experimental measurements that contain only intact ILs are removed. This is done for two reasons; firstly, post burn ILs differ substantially from intact ILs. The inclusion of several neat IL samples in the training data partition may adversely affect the model by allowing features to be selected that will not be present in fire debris samples. Secondly, the focus of this study is to demonstrate results specific to fire debris samples. Due to the use of a Monte Carlo procedure the results presented here are aggregated over 11000 comparisons guaranteeing (by the stratified CV structure) that each of the 110 samples has been classified as part of the testing block 100 times. Hard classification is performed by thresholding the resulting LR value pertaining to the hypotheses stated in Section 2.5. A natural threshold is set at 1 where an LR > 1 provides support in favor of H1 whereas an LR < 1 supports H0 and an LR of exactly 1 means the evidence is equally likely under the competing hypotheses and provides no evidential value. The confusion matrix in Table 2 summarizes this performance. Note that the abbreviations for True Negatives (TN), False Positives (FP), False Negatives (FN), and True Negatives (TN) along with their corresponding true and predicted labels are introduced here.

Accurate classification performance is achieved for an aggregate of 84% of sample comparisons (9269 out of 11000 total comparisons). In addition the low FP rate is a desirable trait for a forensic application. Only samples that are selected as containing ILs are then passed on to the classification step, including both FP and TP samples. This method of reporting results is in line with ILs with the development of an automated method for fire debris pre-screening, wherein samples that are deemed to be ILR negative are unlikely to be examined further. The number of samples that pass from the detection step to the classification step is variable depending on ILR detection performance for that Monte Carlo iteration. The same procedure of stratified cross-validation with 10 data partitions is repeated in another step of class modeling, parameter tuning and classification of test partition samples.

Despite the high specificity observed in Table 2, a fairly low sensitivity can also be noted. In Section 1.1 the labels that serve as ground truth for the evaluation of performance are explicitly defined based on the pre-ignition addition of an IL to a sample. It is possible that for some proportion of the samples, the IL is completely consumed during the course of the fire. This could lead to cases where an sample labeled as containing ILR contains no residual ILR. Previous studies have chosen to define ground truth in terms of labels assigned by an informed analyst (an analyst aware of the presence or absence of a particular IL at the time of sample ignition) [16–18], however in this manuscript we opt to accept the risk of a biased (high) FN rate in order to maintain an absolutely objective labeling scheme, and therefore conservative performance reporting.

3.2. Classification of ignitable liquid residues

Only samples that were predicted to contain ILR are passed on to the classification step, however samples that were incorrectly classified as not containing ILR may yet again be used in the training or parameter tuning partitions. Likewise, the intact IL samples are now allowed in the training and parameter tuning partition in order to encourage features distinctive to the differences between ILR classes to be modeled. The underlying assumption is that a different set of features may be required to distinguish between ILR classes than was required to detect them. A different Monte Carlo simulation is used for the results reported for the classification results reported in Table 3 in order to allow a greater number of iterations and therefore a larger number of samples to pass from the detection step to the classification step. A similar FP rate to the one reported in Table 2 is seen with 80 FP seen out of the 7420 total samples (1.1%) for which classifications were performed.

The overall successful classification of an ILR, assuming a preliminary successful detection step, is 74.9% (5560 of the 7420 classifications performed). The proportion of samples classified in Table 3 will not necessarily reflect the relative class frequencies in the experimental design because there may be a systematic bias in the preliminary detection step (i.e. some ILs are more readily detected and therefore more frequently passed on to the classification step). The results reported for the testing of classification accuracy do not contain any intact liquid samples; only fire debris is classified into the three classes based on the features that are detectable after highly variable fire debris samples have been burnt as per the protocol described in Section 2.1.

It is obvious that the GAS class is distinct within the projected space defined by LDA as quite good classification is achieved for samples containing gasoline ILR, whereas the LO and WS classes are more frequently misclassified as one another, suggesting a possible chemical similarity in these classes.

![Fig. 4. 2-D projection (LDA axes) density maps for the 3 IL classes and new points to be classified.](Image)
An additional observation made when examining specific cases of erroneously classified samples was that occasionally, new samples to be classified would appear very far away from all samples used in the modeling set to define the distributions (as depicted in Fig. 4). This may be an indication that the number of samples used in this study is simply insufficient to achieve better classification results given the complexity of the fire debris matrix used. As with many statistical models relying on the approximation of class characteristics in the form of a probability density function, performance is expected to improve with increased sampling as the

Table 2
Performance observed for the detection of ignitable liquid residue, using the likelihood ratio formulation with a threshold of 1.0.

<table>
<thead>
<tr>
<th>True</th>
<th>ILR (-)</th>
<th>ILR (+)</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.2%</td>
<td>TN = 6692</td>
<td>FP = 57</td>
</tr>
<tr>
<td>39.4%</td>
<td>FN = 1674</td>
<td>TP = 2577</td>
</tr>
</tbody>
</table>
diversity of the variation possible in samples is better captured and
the sample characteristics become more accurately modeled.

Another interesting observation is that the row sums for the
classification data (see Table 3) show a greater tendency for suc-
cessful detection of ILR in samples containing WS compared to
samples containing GAS. This results in a smaller proportion of
GAS samples being passed to the classification stage. This may be
related to the higher volatility of GAS compared to the other ILs,
where perhaps only GAS containing samples with a very strong
ILR compound signal are subject to classification. This observation
can only be made with the prior knowledge that the original data
set is balanced in terms of class representation.

3.3. Interpretation of evidential value

The use of the likelihood ratio framework for the assessment of
evidence allows a direct representation of the projection of new
points into a quantitative assessment of the evidential value. Due
to the dimension reduction performed by LDA, the models produce
a wide range of LRs for a certain type of comparison of samples. For
this reason it is critical to examine the overall behavior of the LR
system. A Tippet plot is generated from the aggregate results
reported in Table 2 and shown in Fig. 6.

A Tippett plot contains information about the proportions of LRs
observed in the system that fall above and below the point of neu-
tral evidence. The x-axis is plotted on a log scale, so 0 is the point at
which evidence is deemed uninformative in supporting either hypothesis. The explanation for the very low observed FP rate is
immediately clear by the portion of the known negative ILR sam-

ers on the right side of the green (dashed) line. While this is a
desirable characteristic for a forensic context, the Tippett plot also
shows that the proportion of false negatives is very high. Subse-
quent classification performance will contain some additional
erroneous classifications of samples that do not contain ILR but
must be nonetheless classified into an ILR type.

We also note that the log10LR values often suggest extremely
large evidential values. Given the size of the dataset and the diffi-
culty of this problem this may be a gross over estimation of the
quantitative evidential value. Tail regions of the modeled distribu-
tions produce values not adequately substantiated by the available
data and their ratios become unstable. More research, including
much more data collection and a full validation of the method
(including calibration of the LR values) is required before the
method can be implemented for casework.

As with the analysis of the ILR detection results, it is of interest
to monitor the behavior of the classification results in terms of the
LRs for competing class membership hypotheses. Performing a
one-versus-one examination allows the generation of a series of
Tippett plots similar to those shown in Fig. 6. In Fig. 7(a), (b), and
(c) the LRs observed for each particular ILR class are shown for
the competing hypotheses in Table 3. In each case two curves sum-
marize correct LRs, that is the LRs for models where the appropriate
class is in the numerator hypothesis ($H_i$). A third line is presented
in each case showing the LRs observed when an ILR positive sam-
ple of a known class is compared against two inappropriate distri-
butions, related to the hypotheses of class membership in the two
remaining classes. These additional lines (plotted in black) should
yield LRs fairly balanced on either side of the neutral-evidence line.

Table 3
Performance observed for the classification of ignitable liquid residues, using the maximum likelihood of class distributions modeled form training samples.

<table>
<thead>
<tr>
<th>True</th>
<th>Predicted</th>
<th>WS</th>
<th>GAS</th>
<th>LO</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILR (−)</td>
<td>16.2%</td>
<td>12.6%</td>
<td>11.0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N = 13</td>
<td>N = 348</td>
<td>N = 303</td>
<td></td>
</tr>
<tr>
<td>LO</td>
<td>3.9%</td>
<td>75.4%</td>
<td>12.6%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N = 81</td>
<td>N = 2109</td>
<td>N = 303</td>
<td></td>
</tr>
<tr>
<td>GAS</td>
<td>12.6%</td>
<td>75.4%</td>
<td>11.0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N = 789</td>
<td>N = 2109</td>
<td>N = 303</td>
<td></td>
</tr>
<tr>
<td>WS</td>
<td>11.0%</td>
<td>12.6%</td>
<td>11.0%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N = 57</td>
<td>N = 55</td>
<td>N = 1431</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 6. Tippett plot of likelihood ratios observed for ignitable liquid detection.
These plots provide additional information pertaining to the magnitudes of LRs observed and the cumulative proportion of LRs falling in that range for correct and erroneous evidential assertions.

The performance reported in Table 3 is well explained by the LRs observed. Both sets of hypotheses for the GAS result in a high proportion of LRs on the right side of the neutral evidence line in Fig. 7(b). Samples from the WS class show similar LRs for both denominator hypotheses and a proportion of LRs on either side of the neutral evidence line as would be expected from the reported classification performance. Interesting behavior is observed for the LO samples, which are well differentiated from GAS samples, obvious from the red line in Fig. 7(b). But this is not the case when discriminating LO samples from WS samples, as seen by the blue line in Fig. 7(c).

4. Discussion

4.1. Compounds of interest detected

Over the course of many iterations with intermediate randomization of sample order, it is possible to keep track of which
features in the RFR are most frequently selected. As mentioned in Section 2.4, the feature selection retains full resolution in terms of the nominal mass values and retains some spatial resolution (within the tile region).

This allows us to depict the original chromatographic data in terms of a Extracted Ion Chromatogram (EIC) wherein each tile region is plotted using only the ions selected as useful for that region. This figure is useful in exploring the identity of compounds that are selected as useful in detection and classification of fire debris samples based on their ILR content. Figure 9(b) shows the result of such a plot. This visual tool is essential in forensic casework in order for the forensic chemist to understand the output of the numerical methods and likewise for the statistical understanding of the chemical basis of evidential assertions.

One of the peaks observed in Fig. 8(b) is not even visible in the corresponding location of Fig. 8(a). This compound was present in very low abundance in this sample. Its apparent absence in the TIC is due to the color scaling required to visualize the range of intensities in either plot. The mass spectra of the peak can however be isolated and are shown in the lower left hand corner of Figure (a). Based on the mass spectrum this compound is tentatively identified as a C3 substituted benzene (nominal molecular weight 120, molecular formula C_9H_{12}). Compounds of this class are expected to be present in most ILs.

The EIC’s generated for several samples were used to retrieve the LIS which could then be passed as query spectra to the NIST/EPA/NIH Mass Spectral Database (NIST 11). For each query spectra the top 3 candidate compounds returned as library hits were manually examined. To ensure confident identification all identified compounds were searched for in the main library, using the NIST MS Search program (NIST MS Search 2.0), the returned spectra were manually compared to the LIS spectra suspected to contain the candidate compound. Some example compound matches and corresponding reference spectra from the NIST library can be seen in the Supplementary Material.

Systematic examination of compounds regularly discovered in samples known to have contained particular ignitable liquids in the experimental design allows the identification of compounds indicative of the presence of particular ignitable liquid classes. Table 4 summarizes several compounds that were repeatedly determined useful in the classification of samples into different IL categories.

4.2. Future work

Preliminary results have been presented showing that a two-dimensional chromatographic separation allows the isolation of valuable features for the classification and interpretation of fire debris samples without having to directly address complications related to peak integration. Current results are based on a coarse
division of the chromatographic space into 12 tiles. An immediate consideration for future work is the exploration of different strategies for dividing the chromatographic space and computing LIS features. Intuitively, the regions should be as small as possible to include the least amount of interfering noise and the greatest proportion of interesting analytes for the forensic problem. Conversely, shrinking the tiles approaches a pixel by pixel comparison which introduces problems with inter-sample alignment. This is the objective of ongoing research.

In Section 4.1 an example is presented demonstrating the potential assistance in identifying chemical characteristics of forensic interest for the detection and classification of ILR in fire debris. The result is achieved by visualizing the original analytical data while applying different ion filters over the spatial regions corresponding to the original division of the chromatogram. The results presented here demonstrate the potential utility of this strategy in assisting a forensic chemist with the examination of a case sample. Currently underway is an extensive survey aimed at providing an extensive analysis of chemical markers of potential interest in ILR detection. This will be accomplished by using the procedure depicted in Fig. 8(b) to search for additional trace compounds of interest in other samples and determine if new indicators of ILR can be discovered. Of additional interest would be the future exploration of the evidential value of new markers in a targeted approach for ILR detection and classification.

5. Conclusions

A novel method for performing detection and subsequent classification of ILR has been presented using GC × GC–MS analyses of fire debris samples. Preliminary findings demonstrate reasonable detection rates (84%) with low FP occurrence (<1%) on a set of realistic fire debris samples. Current classification rates present a realistic prospective on the applicability of this method to automated prescreening of fire debris samples in a forensic context.

A two-step evidence-based feature selection method is presented for GC × GC–MS chromatograms, which demonstrates the potential of this analytical technology for use in fire debris forensics. The reduction from a very large dimensionality to a robust and information rich reduced feature space is demonstrated. This method takes advantage of the vast separation space while avoiding the shortcomings typically encountered in GC × GC. The feature selection protocol provides intuitively interpretable selection and discovery of pertinent chemical information, isolating compounds in both the time domain and in the mass domain. This allows experts to pinpoint the source of the characteristics in the original chemical analysis that are the main basis for the evidential assertions reported. By projection of the already reduced feature set into a suitable space for classification via LDA, the desired size of the reduced feature space itself is flexible and should theoretically yield better performance with the availability of larger forensic reference collections.

The method also allows the calculation of, in a fully objective manner, the evidential values (LIs) for comparing two hypotheses of interest. Although promising results were obtained, further research and extensive additional data collection is necessary for casework implementation.

Acknowledgement

This research was funded by NWO Grant No. 727.011.006, under the research objectives of the COMFOR project. The authors acknowledge and appreciate the contribution of Brenda van Daelem in providing laboratory assistance during sample preparation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.forc.2016.10.003.

References

[26] Kari L. Organtini, Anne L. Myers, Kari J. Jobst, Jack Cochran, Brian Ross, Brian McCarty, Eric J. Reiner, Frank L. Dorman, Comprehensive characterization of the
halogenated dibenzo-p-dioxin and dibenzofuran contents of residential fire
debris using comprehensive two-dimensional gas chromatography coupled to

Heather D. Bean, Jane E. Hill, Jean-Marie D. Dimandja, Improving the quality of
biomarker candidates in untargeted metabolomics via peak table-based
alignment of comprehensive two-dimensional gas chromatography-mass

Hadi Parastar, Mehdi Jalali-Heravi, and Roma Tauler, Comprehensive two-
dimensional gas chromatography (GC × GC) retention time shift correction
and modeling using bilinear peak alignment, correlation optimized shifting
and multivariate curve resolution. Chemometrics and Intelligent Laboratory
African-European Conference on Chemometrics, Rabat, Morocco, September
2010 Special Issue Section: Preprocessing methods Special Issue Section:
Spectroscopic imaging.

Martin Lopatka, Andrei Barcaru, Marjan J. Sjerps, Gabriel Vivó-Truyols,
Leveraging probabilistic peak detection to estimate baseline drift in complex

M. Adahchour, J. Beens, U.A.Th. Brinkman, Recent developments in the
application of comprehensive two-dimensional gas chromatography, J.
Chromatography.

Jens Dallüge, Jan Beens, Udo A.Th Brinkman, Comprehensive two-dimensional
gas chromatography: a powerful and versatile analytical tool, J. Chromatogr. A

Anders Nordgaard, Birgitta Rasmusson, The likelihood ratio as value of evidence

G. Zadora, A. Martyna, D. Ramos, C. Aitken, Statistical Analysis in Forensic

Annabel Bolck, Haifang Ni, Martin Lopatka, Evaluating score- and feature-
based likelihood ratio models for multivariate continuous data: applied to