

## Evaluation of data preprocessings for the comparison of GC-MS chemical profiles of seized cannabis samples

Slosse, A; Van Durme, F; Samyn, N; Mangelings, D; Vander Heyden, Y

*Published in:*  
Forensic Science International

*DOI:*  
[10.1016/j.forsciint.2020.110228](https://doi.org/10.1016/j.forsciint.2020.110228)

*Publication date:*  
2020

*License:*  
CC BY-NC-ND

*Document Version:*  
Accepted author manuscript

[Link to publication](#)

*Citation for published version (APA):*  
Slosse, A., Van Durme, F., Samyn, N., Mangelings, D., & Vander Heyden, Y. (2020). Evaluation of data preprocessings for the comparison of GC-MS chemical profiles of seized cannabis samples. *Forensic Science International*, 310, [110228]. <https://doi.org/10.1016/j.forsciint.2020.110228>

### Copyright

No part of this publication may be reproduced or transmitted in any form, without the prior written permission of the author(s) or other rights holders to whom publication rights have been transferred, unless permitted by a license attached to the publication (a Creative Commons license or other), or unless exceptions to copyright law apply.

### Take down policy

If you believe that this document infringes your copyright or other rights, please contact [openaccess@vub.be](mailto:openaccess@vub.be), with details of the nature of the infringement. We will investigate the claim and if justified, we will take the appropriate steps.

# **Evaluation of data preprocessings for the comparison of GC-MS chemical profiles of seized cannabis samples**

A. Slosse<sup>1,2</sup>, F. Van Durme<sup>2</sup>, N. Samyn<sup>2</sup>, D. Mangelings<sup>1</sup>, Y. Vander Heyden<sup>1,\*</sup>

<sup>1</sup> Department of Analytical Chemistry, Applied Chemometrics and Molecular Modelling, Vrije Universiteit Brussel (VUB), Laarbeeklaan 103, B-1090 Brussels, Belgium

<sup>2</sup> Department Drugs and Toxicology, National Institute for Criminalistics and Criminology (NICC), Vilvoordsesteenweg 100, B-1120 Brussels, Belgium

\* Corresponding author: Tel.: +32 2 477 47 34

Fax: +32 2 477 47 35

Email: [yvanvdh@vub.be](mailto:yvanvdh@vub.be) (Y. Vander Heyden)

## **Abstract**

Cannabis is the most frequently used illicit drug in Belgium, where it is mainly cultivated indoor. To improve the fight against this drug, cannabis-profiling methods are required. Cannabis is a natural product and its chemical composition depends on many factors, which cause a high heterogeneity and variability in the secondary metabolites, and make this study challenging. The aim of this study is to combine cannabis profiling with statistical methodology to evaluate the intra (within)- and inter (between)-plantation variabilities with the goal to define a suitable approach linking seized marijuana to given plantations. The data set used contains 46 samples from 9 locations. The chemical profiles, consisting of data from eight cannabinoids, are obtained by gas chromatography - mass spectrometry. The raw data (peak areas) is pretreated with different preprocessing methods. The Pearson correlation coefficients between intra-location profiles were calculated after each pre-treatment, and the 95 and 99% confidence limits determined. All preprocessed data were then compared with the internal standard normalization reference method with the aim to minimize the overlap between intra- and inter-location results, i.e. to reduce the number of false positives, and to obtain the best discrimination. Furthermore, cross-validation was used to evaluate the model originating from the most efficient data pre-treatment technique. The best results were obtained, when the peak areas were normalized to the internal standard with subsequent calculation of the fourth root. It results in a reduction of false positives for both confidence limits to 11% and 14% compared to 21% and 27% for the reference method. Cross-validation reveals similar false positive results as for the calibration set. In conclusion, when preprocessing the data, an improved model is obtained resulting in a significant decrease in the number of false positives. After studying the predictive performance of the model, it appears to be representative for the entire plantation information.

**Keywords** Cannabis profiling, intra- and inter-location variabilities, data preprocessing methods, Pearson correlation classification, cross-validation.

## 1. Introduction

*Cannabis sativa* L. belongs to the *Cannabaceae* family and is well known for its illegal abuse (1–4). This psychoactive plant remains nowadays worldwide the most commonly used illicit drug. The 2018 World Drug Report revealed that 4.7 tons of herbal cannabis, i.e. marijuana, had been seized worldwide and 192.2 million persons used it (5, 6).

Cannabis misuse has also been observed in Belgium, where 26.587 seizures occurred in 2016 (7). In this country, cannabis is mainly cultivated indoors because of its cold climate. Doing so, cultivators also have the advantage to remain hidden for law enforcement (8–11). Most important, conditions, such as temperature, humidity, light and nutrient supply, can be optimized to obtain maximal yield of the cannabis crops with multiple growths annually (4, 12–15). Mainly female flowering tops are used for consumption because of their highest concentration in  $\Delta^9$ -tetrahydrocannabinol ( $\Delta^9$ -THC), the main component causing the mind-altering effects of the drug (16, 17).

Unpublished data from the Belgian federal police show a large increase in the number of confiscated cultivation sites in recent years (Fig. 1). In 2007, 398 cannabis plantations were discovered ranging from a micro level, containing 2 to 5 plants, up to an industrial level, containing more than 1000. By 2017, the number of seizures had tripled for almost all plantation sizes.

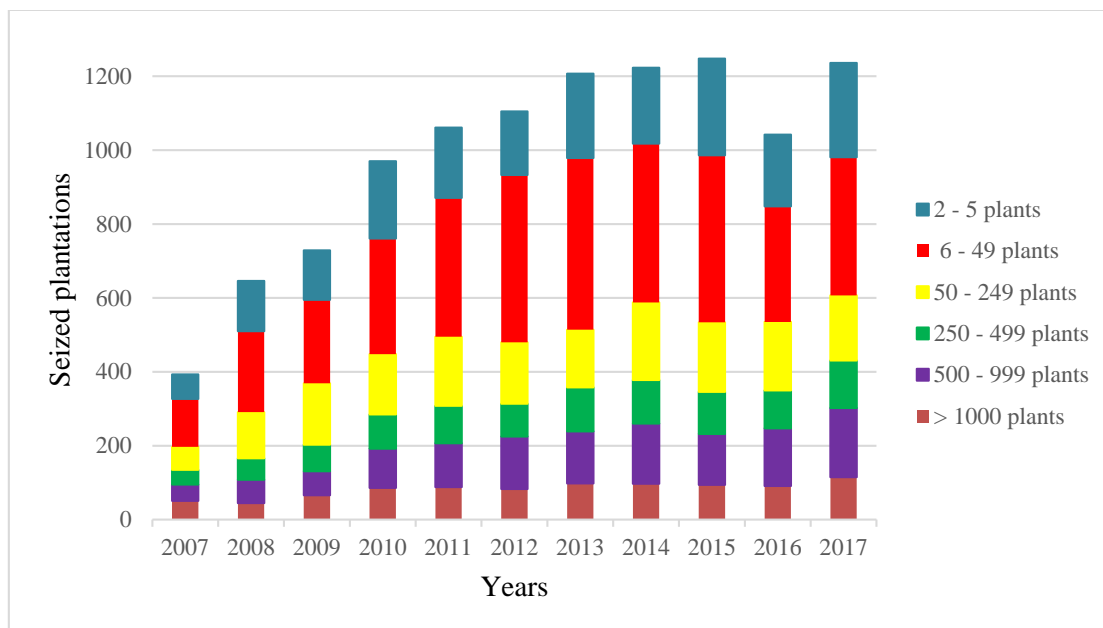


Fig. 1. Number of seized cannabis plantations in Belgium, distinguishing 6 different plantation sizes, over a ten year time period (2007-2017) (Unpublished Belgian federal police data).

In view of prosecution, magistrates often raise questions with respect to cannabis-related crimes. For instance, when marijuana has been seized from two dealers and/or consumers and law enforcement has already collected different pieces of information between both seizures such as telephone data, that could lead to a possible connection, would it be possible that a forensic expert is able to determine a common source or history between the two persons and confirm whether there is a link between the seizures by analyzing both cannabis seizures? To answer this question, cannabis profiling can be a very useful tool and provide complementary information to that already gained by police investigations (18, 19).

In general, drug profiling can be described as the expression of physical and/or chemical characteristics of a drug, which are chosen in relation to the specific purposes of the analytical analysis. Consequently, these profiles are used for different policies such as law enforcement, legislation and public health (20, 21). For illicit drugs, profiling concerns the application of methods to determine chemical substances and their quantities and/or physical features with respect to drug seizures. The profiles specifically are used to compare seizures for

statistical/tactical intelligence- or evidence-based objectives (20, 22, 23). Illicit-drug profiling aims to gather information from drug samples to find a link between drug seizures and cultivations (common origin), manufacturing processes (synthetic production), precursors, drug trends trades and distribution (dealer-user networks), or supply sources. The intention herewith is to support drug enforcement agencies in their battle against organized drug crime, mainly in view of the identification of drug trafficking organizations and the subsequent disruption- and abolition of their criminal activities (18, 23, 24).

Cocaine, amphetamine and heroin profiling has already been demonstrated of utmost importance in forensic science to establish links between seizures and/or samples (25–27). Andersson et al. (27) and Locicero et al. (26) successfully demonstrated the use of correlation and distance parameters in combination with several data pretreatments on the raw data profiles to differentiate between linked and non-linked amphetamine and cocaine samples (26, 27). The latter group also applied statistical confidence intervals to estimate the discriminating power of the used method (26).

Earlier work on cannabis extracts has already been done for other forensic applications, and gas chromatography was the most frequently used separation technique (17, 28–30). Classification of seized cannabis into fiber- or drug-phenotypes was performed for juridical purposes (31, 32). The estimation of the age of seized cannabis samples was studied using two cannabinoids, i.e.  $\Delta^9$ -THC and its degradation product cannabitol (CBN) (16, 33). The determination of the yield of marijuana from indoor cultivation, taking several growth parameters into account, was carried out to penalize the perpetrator by estimating the gained profits of the cannabis crops (14, 34). Stefanidou et al. (35) attempted to use the three main components, i.e.  $\Delta^9$ -THC, CBN and cannabidiol (CBD), to distinguish samples originating from outdoor cultivation in different geographical regions, but no satisfactory method for adequate distinction was found. Afterwards, multivariate data analysis and chemometric techniques, such as principal

component analysis, hierarchical cluster analysis and classification methods, were introduced as new approaches but mainly focused on the classification of herbal cannabis (36).

Only very few publications are available with regard to the actual direct comparison of seized cannabis samples. This controversial plant is a natural product and its chemical composition, as in other herbal material, depends on many factors, e.g. light, temperature, fertilization, and storage, which cause a heterogeneity in the secondary metabolites (37, 38). Finding a suitable method to compare different seizures is therefore challenging. In this study, the authors knowledge on cannabis profiling for forensic objectives, where there is a need for comparison of different samples, is based on very limited published data.

The aim of this study is to combine cannabis profiling with multivariate data analysis and statistical methodology, focusing on the cannabinoid contents obtained with GC-MS, to evaluate the intra (within)- and inter (between)–seized plantation variabilities. Different data pre-processing methods will be compared on their capability to resolve between samples from the same or different cultivation sites, with as goal to properly define an acceptable threshold value to link seized marijuana samples.

## **2. Experimental**

### **2.1. Seized cannabis samples**

Forty-seven marijuana samples were collected from indoor cultivation rooms in 9 regions of Belgium and analyzed using GC-MS. Each sample includes the mature flowering tops originating from one single plant. A comprehensive overview of the data set and the different seizures can be found in table 1.

Table 1. Number of cannabis plants per confiscated plantation (location).

Location	Number of samples (plants)
1	5
2	5
3	10
4	2
5	2
6	10
7	5
8	2
9	6
<b>TOTAL</b>	<b>47</b>

## 2.2. Standards and chemicals

Tribenzylamine (TBA) purchased from Alfa Aesar<sup>®</sup> (Karlsruhe, Germany) is used as internal standard. Denaturated ethanol (99% v/v), for the extraction of the herbal material, is obtained from VWR BDH Prolabo Chemicals<sup>®</sup> (Leuven, Belgium).

## 2.3. Sample preparation

The flowering buds were dried in an oven at 40°C for at least 12 hours. The herbal material was grinded using an IKA<sup>®</sup> Tube Mill Control (IKA, Stauffer, Germany) and afterwards homogenized and accurately weighed ( $100 \pm 10$  mg) into 12-ml glass test tubes. The weighed samples were subjected to the extraction step, applying the Tecan Freedom Evo 150 liquid handling platform (Tecan Trading AG, Männedorf, Switzerland). This instrument adds 10.0 ml ethanol 99%, containing 0.01 mg/ml TBA to each test tube. Then, all samples were transferred into an ultrasonic bath for 15 min and horizontally shaken for 5 min (100 turns). Subsequently, 1.0 ml of the obtained solution was transferred into glass vials, sealed and subjected to chromatographic analysis.



## 2.4. GC-MS settings

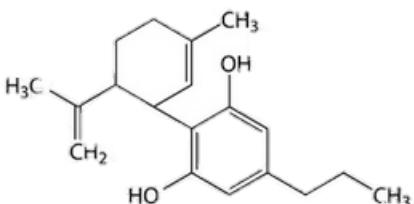
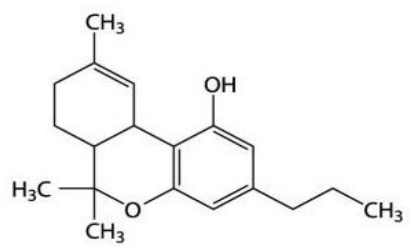
The experiments were performed on an Agilent® 6890N gas chromatographic system (Agilent Technologies, Santa Clara, CA, USA) coupled to an Agilent® 5973N mass selective detector. Separation was achieved on a DB-5ms (5% diphenyl, 95% dimethyl silica) capillary column (15 m x 0.25 mm i.d., 0.25 µm film thickness, J&W Scientific, Folsom, CA, USA) with helium as carrier gas at a constant flow rate of 1.3 mL/min. The injections were carried out using an Agilent 7683 Series injector. The injector temperature was 230°C. The injection volume was 2.0 µl and the split flow was set at 10:1. The oven temperature started at 60°C, increased at 8.5°C/min to 240°C and was held for 2 min. The total run time was 23 min. The MS analyses were performed in full-scan mode with masses in the 50-500 Da range. The temperature of the ion source and the single quadrupole analyzer were 300°C and 150°C, respectively.

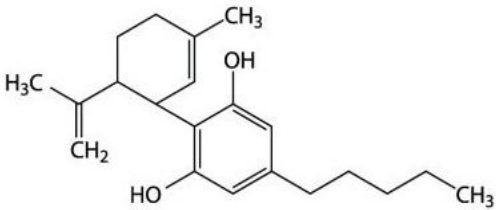
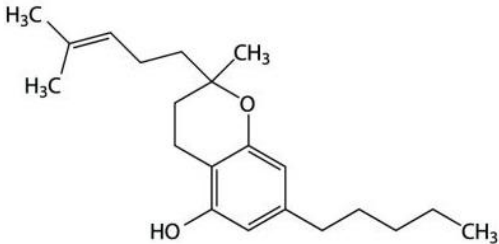
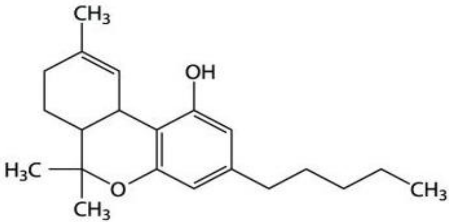
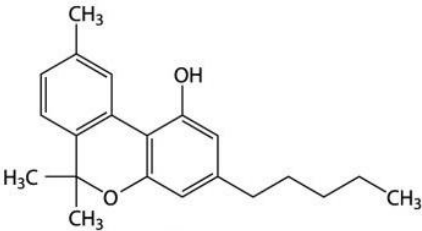
## 2.5. Cannabinoids

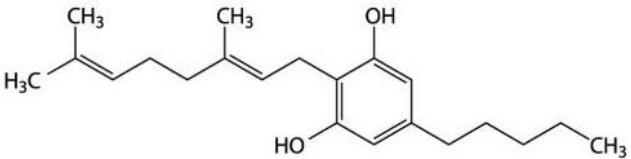
As in Andersson et al. (27), the target cannabinoids were chosen when they could be easily identified, when repeatable peaks were found and when they were stable over time. In this study, a reference chromatogram, originating from freshly seized marijuana, was used to determine the compounds of interest. Nine cannabinoids were visible and selected as variables for the chemical profile, i.e. cannabidivarin (CBDV),  $\Delta^9$ -tetrahydrocannabivarin (THCV), cannabichromene (CBC), cannabidiol (CBD), CBN, cannabigerol (CBG),  $\Delta^9$ -THC, and unknown compounds CB1 & CB2. CB1 was chosen based on the paper by Fishedick et al. (39). CB2 was another unknown cannabinoid that was selected with its characteristic mass ions. To characterize the peaks in the chromatogram, target and qualifier ions were obtained from reference standards, published articles and commercial databases. Subsequently, injection and extraction replicates were used to evaluate the analytical method performance and the variability within a plant. Intra- and inter-plantation variabilities were also investigated for the selected cannabinoids (unpublished results). A summary of the compounds can be found in

table 2. The databases used are the National Institute of Standards and Technology standard reference database (United States Department of Commerce, version 2.0d, 2005, Gaithersburg, Maryland, USA), mass spectral library of drugs, poisons, pesticides, pollutants and their metabolites database (Wiley-VCH, 1<sup>st</sup> edition, 2011, Weinheim, Germany) and mass spectra of designer drugs database (Wiley-VCH, 2014, Weinheim, Germany). The target compounds were further studied with respect to their availability in other samples.

Table 2. Retention times ( $t_R$ ), cannabinoid structures (when compound is known), target ion  $m/z$  and qualifier ions  $m/z$  of all compounds of interest in the extract.

$t_R$ (min)	Structure compound	Target ion ( $m/z$ )	Qualifier ions ( $m/z$ )
18.01	<b>Cannabidivarin (CBDV)</b> 	203	174 121 286
19.00	<b><math>\Delta^9</math>-Tetrahydrocannabivarin (THCV)</b> 	203	243 271 286
20.03	<b>Cannabidiol (CBD)</b>	246	314 271 299

			
20.20	<b>Cannabichromene (CBC)</b> 	231	174 299 314
20.50	<b>CB1</b> <b>Unknown</b>	297	313 356 243
20.63	<b>CB2</b> <b>Unknown</b>	231	299 314
21.10	<b><math>\Delta^9</math>-Tetrahydrocannabinol (<math>\Delta^9</math>-THC)</b> 	243	299 231 314
21.54	<b>Cannabinol (CBN)</b> 	295	296 238 310
21.57	<b>Cannabigerol (CBG)</b>	193	231

			316 123
--	--	--	------------

## 2.6. Data analysis

MSD Enhanced Chemstation F.01.03.2357 (Agilent Technologies, Santa Clara, CA, USA) was used as software to process all chemical profiles and to integrate each target peak in the extracted ion chromatogram. The area under curve (AUC), as chromatographic response, is recorded for all selected peaks. Calculations were carried out by means of Microsoft® Excel 2013 (Microsoft Corporation, Redmond, WA) software and Matlab™ 9.4 (The Mathworks, Natick, MA). Data pre-processing methods were also executed using both above-mentioned programs. The percentages false negatives (FN) and false positives (FP) were calculated in Excel. For the Pearson correlation coefficient computations and the determination of the threshold values at the 95% and 99% confidence levels, m-files were used, written in Matlab version 9.4. IBM SPSS Statistics version 26 (IBM, Armonk, NY, USA) was used to draw the ROC curves.

### 2.6.1 Data pretreatment on the chemical profiles

To increase and optimize the discriminating power between the cannabis samples from different cultivation sites, a number of pretreatment methods are tested on the AUCs of the chemical profiles. Preprocessing is often done before multivariate data analysis to create a better distinction (40). It is needed to reduce instrumental influence, redundant variation and to decrease the large concentration differences between the cannabinoids. Consequently, a better comparison can be obtained with less influence of the amount of main common components (41–43). Several transformation techniques were presented on the raw data, including the internal standard normalization, where each value is divided by the area of the internal standard

(44), row scaling to constant total (each value in the profile is divided by the total sum of all peak areas per chromatogram) (44, 45), normalization using logarithm or square/fourth root (27, 46), standardization (each column value divided by its respective variable standard deviation) (27), auto-scaling (use of mean centering and variance scaling) (47), minimum-maximum normalization (linear transformation to values between 0 and 1) (48), and some combinations of the above-mentioned methods.

### 2.6.2 Similarity analysis – Pearson correlation coefficient

Pearson correlation is a numerical method that compares continuous pairs of variables/samples/profiles (27). In fact, it assesses the possible strength and direction of any linear relationship between these variable pairs. The Pearson correlation coefficient (PCC), denoted as ‘r’, belongs to the most used statistical metrics within, among other, data analysis, decision making, forensic and biological research (49–53). The PCC ( $r_{(x_i, y_i)}$ ) of the variables  $x_i$  and  $y_i$  is calculated as (Equation 1):

$$r_{(x_i, y_i)} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

Where  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ ,  $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$ , n the number of variables per profile and  $x_i$  and  $y_i$  the equivalent values in two profiles.

PCC has already been efficiently used in cocaine and amphetamine analysis, with regard to optimal distinction of linked and non-linked samples of seized drugs (26, 27, 54). It is the intention to use also the correlation measurement in this cannabis research to pairwise compare the chemical profiles, within and between the seized cannabis samples, obtained from the 9 plantations.

## **2.7. False negative and false positive errors**

When comparing samples, different outcomes can be seen depending on the used pretreatment method. Fig. 2A shows the perfect separation between the intra-location and inter-location samples. Here, all linked samples are determined as 100% true positive (TP) and non-linked samples 100% true negative (TN), for which a link can be confirmed with the highest certainty. However, in drug profiling, the occurrence of an overlap between the distributions is often observed (Fig. 2B). These FNs and FPs are used for evaluating the discriminating power of the preprocessing method. To answer the question of the magistrate, whether two confiscations are linked, there is a tendency to minimize the FPs to be sure that the different marijuana samples are related or not. However, this is accompanied by a risk to increase the rate of FNs, resulting in a certain number of linked samples that are seen as unlinked. Therefore, a proper balance between both parameters is of interest (52, 55, 56).

In forensic science, statistical methodology to analyze and interpret the data is often applied. In combination with correlation calculations, a threshold value can be defined. This value is very important for forensic analysts because it acts as a decision-maker whether two samples are linked or not. In order to define the threshold, confidence limits of the preprocessed data are determined in an attempt to find an acceptable FP rate (46, 57).

Confidence limits for the correlation coefficient calculations were estimated at the traditional 95% and 99% levels for normally distributed data (43, 58, 59). Similar approaches were already applied earlier in illicit drug profiling, for instance, on heroin (46) or cocaine (60) results.

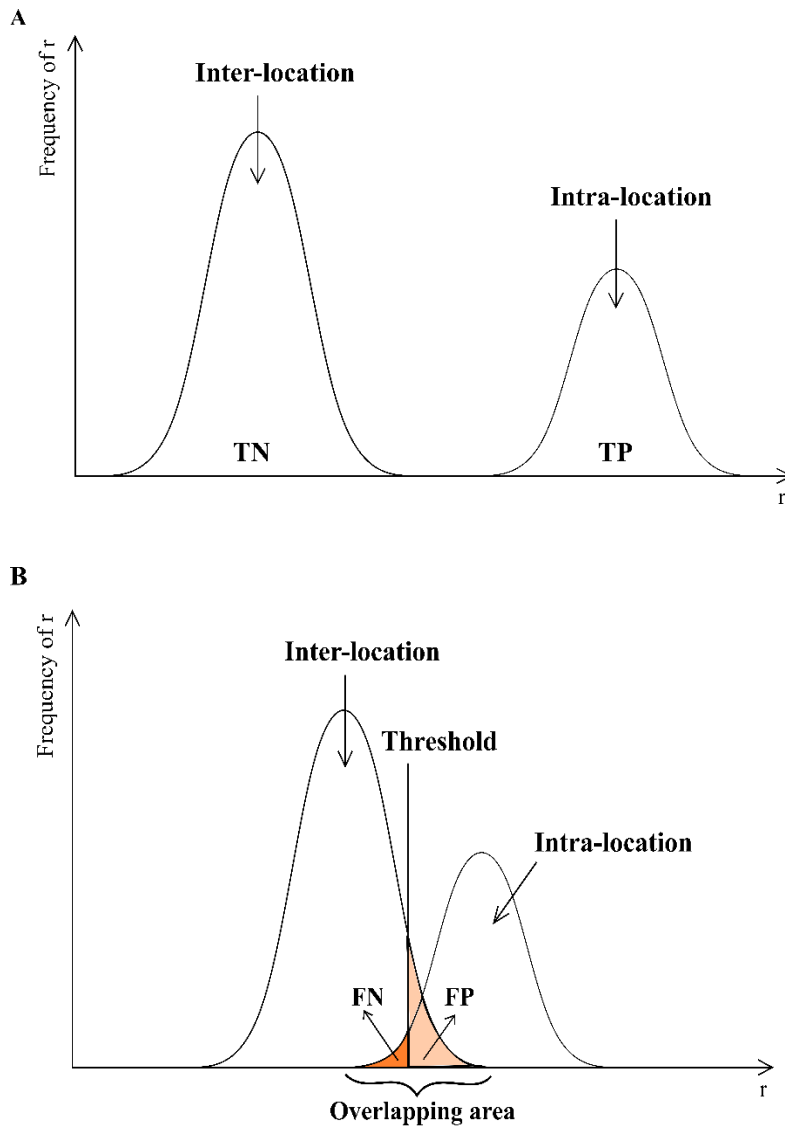


Fig. 2. (A). Perfect situation showing linked and non-linked samples without overlap in  $r$  and (B) Overlap of the  $r$ -values resulting in FN and FP errors. FN = percentage linked  $r$ -values considered as unlinked. FP = percentage unlinked  $r$ -values seen as connected.

### 2.7.1 Roc curves

Receiver operating characteristics (ROC) analysis is a tool to study the accuracy of the pretreatment to discriminate the intra- and inter-location samples (26). Drawing the ROC curve consists of plotting the sensitivity (the TP rate) as a function of the 1-specificity (the FP rate) for various thresholds. The area under the ROC curve (AUC) is a measure for the distinguishing

performance of the pretreatment method. This area can vary from 0.5, i.e. total overlap of the intra- and inter-location samples, to 1 which represents a perfect discrimination of the samples. Consequently, the faster the increase in the ROC space, the more accurate the distinction is between linked and non-linked samples (61, 62).

### 3. Results and discussion

#### 3.1. Cannabinoid profile

After sample preparation and GC-MS analysis, the cannabinoids are identified as already presented in table 2. The typical chromatogram of a sample is shown in Fig. 3. As can be observed in Fig. 3, all samples are strongly dominated by  $\Delta^9$ -THC. The other components occur in small amounts compared to the main peak.

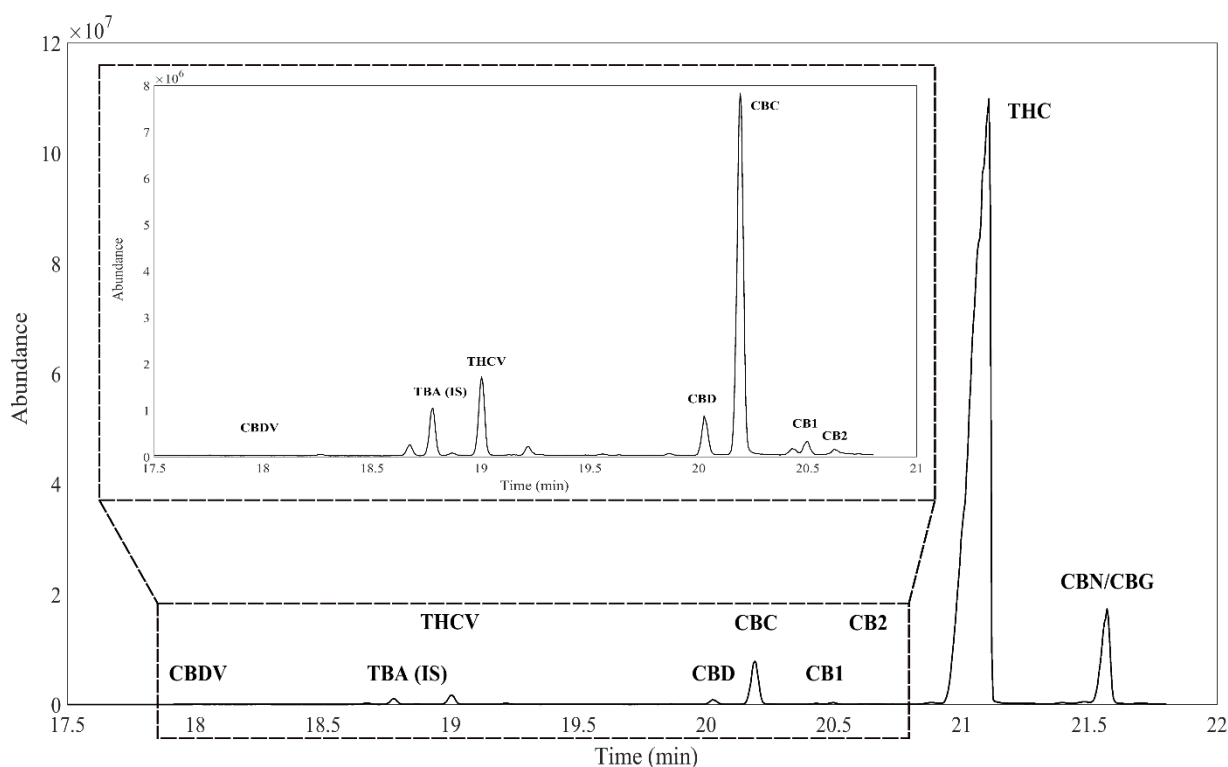


Figure 3. Part of the GC-MS chromatogram of a marijuana sample seized from location two. Nine peaks are initially used, corresponding to nine cannabinoids. Peak labels as in table 2.

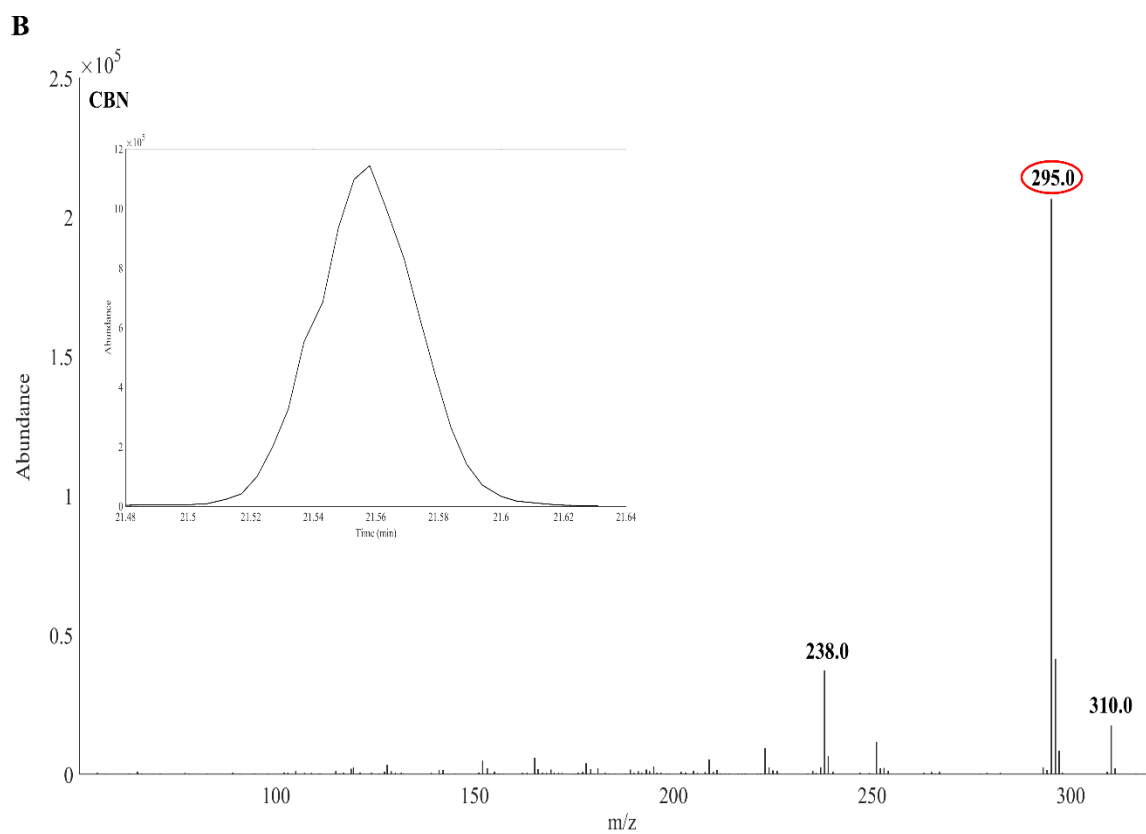
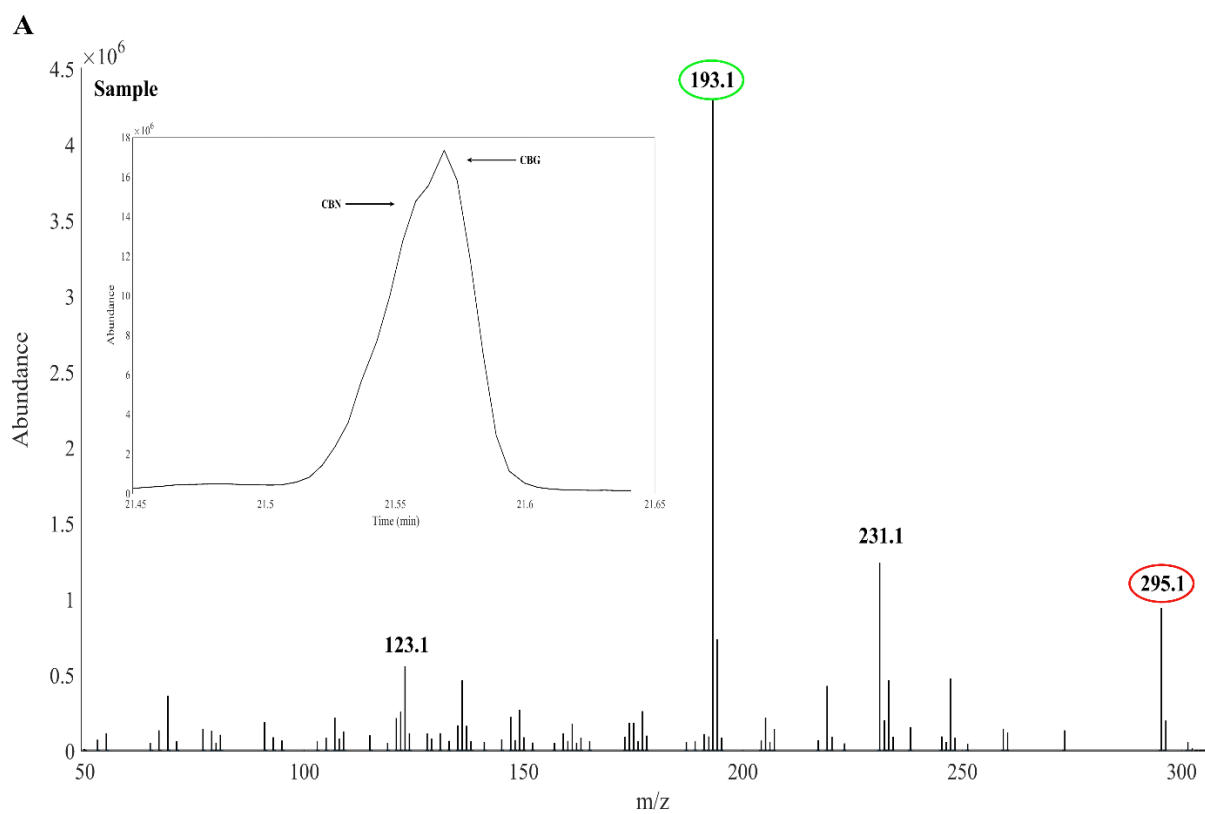


Quantification of  $\Delta^9$ -THC, CBN and CBD was carried out, with the validated routine analysis NICC method (GC-FID technique) to determine the potency of indoor cultivated cannabis. According to the Belgian legislation, herbal cannabis with a concentration higher than 0.2 % of  $\Delta^9$ -THC and tetrahydrocannabinolic acid, is classified as drug-type cannabis. The  $\Delta^9$ -THC content ranged from 13 % to 20 % for all locations, while CBD and CBN were below 1 % (Table 3). The low percentage of CBN confirms that fresh herbal material is being studied because this component is the oxidation product of  $\Delta^9$ -THC and indicates the age of marijuana.

Table 3. Average content in concentration m/m% of the three main components in herbal cannabis. Number of samples per location: see Table 1.

Location	1	2	3	4	5	6	7	8	9
Cannabinoids									
$\Delta^9$ -THC	14.2	19.4	19.4	13.4	15.2	17.1	20.2	16.0	15.3
CBD	0.22	0.55	0.39	0.31	0.43	0.73	0.66	0.37	0.41
CBN	0.11	0.08	0.09	0.07	0.18	0.05	0.07	0.19	0.07

In the total ion chromatogram (Fig. 3), all peaks are baseline separated except for CBG and CBN. When analyzing complex samples with GC-MS, it is common knowledge that co-elution may occur (63). Both peak areas could still be determined by focusing on the different m/z-values of each component, allowing the two signals to be separately detected in the overlapping peak. As presented in figure 4, from the mass spectrum of a sample, both components can be found and quantified. The reference spectrum of CBN shows that the target ion 295.0 is absent in the CBG mass spectrum, while that of CBG (m/z-value 193.0) is not present in the mass spectrum of CBN as well. In other words, the two compounds have an exclusive ion, resulting in two mutually unbiased areas. The chosen m/z-ratios can be found in table 2.



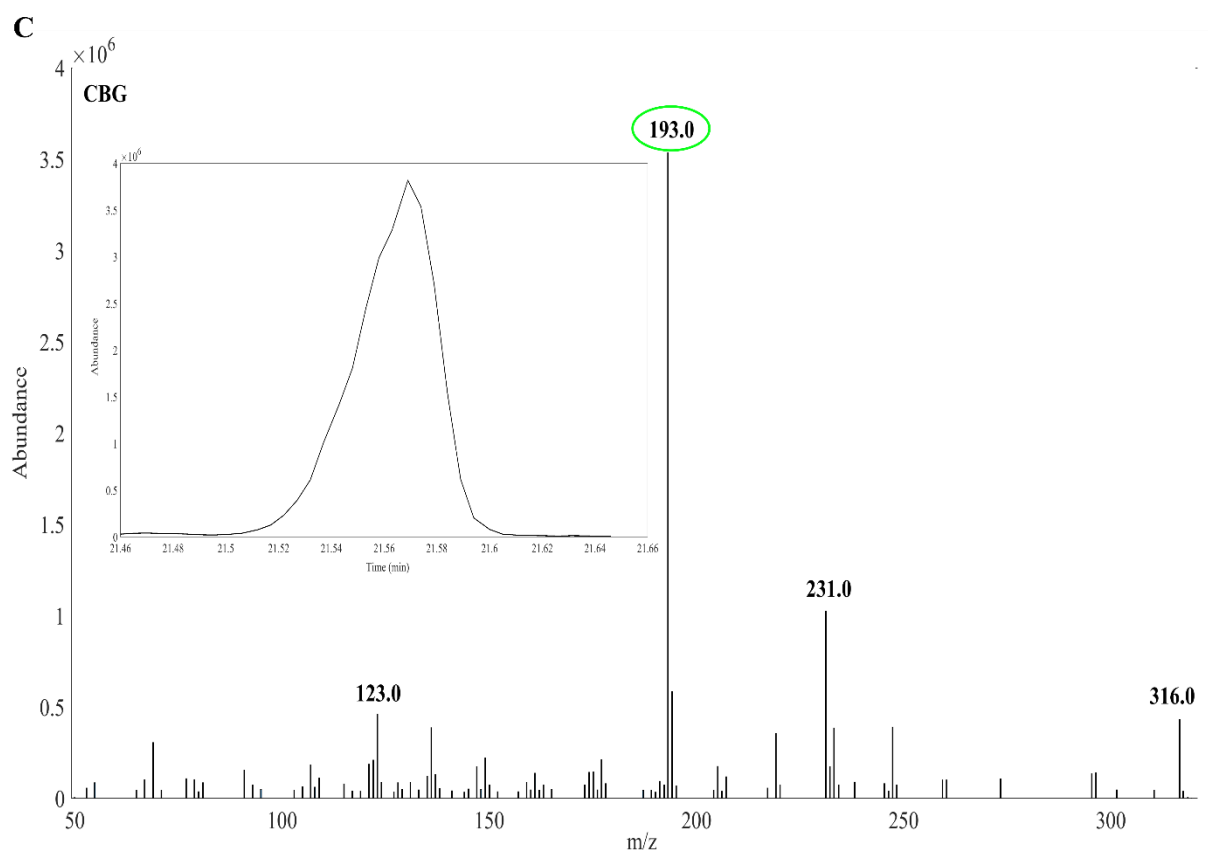


Fig. 4.(A) TIC signal (insert) and the corresponding mass spectrum, showing the co-elution of CBN and CBG. (B and C) Mass spectra of CBN and CBG, respectively, with their unique  $m/z$ -value as target ion, indicated by the green and red colored circles (Inserts: extracted target ion chromatographic signal of both cannabinoids).

One substance, i.e. CBDV, was removed from the profile because this component was most difficult to identify and quantify. Very small peak areas were detected in comparison with the other cannabinoids and often no peak could be integrated (Fig. 5). Zero values (for instance, used when compound is not detected) have an influence on the Pearson correlation coefficient between variables and affect the multivariate data analysis (64). Thus, CBDV may cause erroneous data variability between plants from the same cultivation site and was consequently removed.

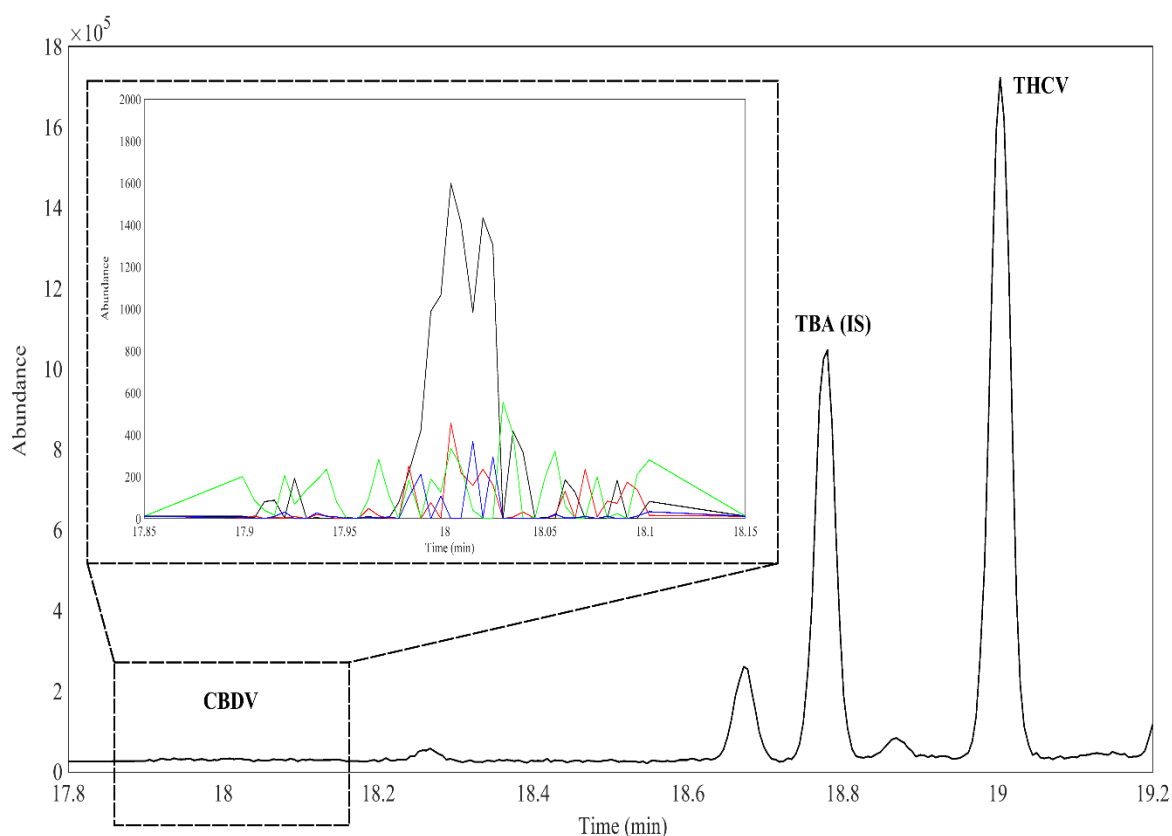


Fig. 5. TIC and extracted ion chromatogram (insert) for CBDV. Black, red, green and blue lines in the insert figure represent the unfound target or qualifier ion.

### 3.2. Data preprocessing

Because of the dominance of  $\Delta^9$ -THC in the chemical profile, several preprocessing techniques were applied on the data set in order to reduce its influence and to increase the relative impact of lower concentrated compounds in order to obtain a better distinction between the samples.

The following pre-treatments and their combinations were studied:

Normalization by IS peak (N), N followed by square root, N followed by fourth root, N followed by a logarithm transformation, N followed by row scaling to a constant total, N followed by standardization (S), N followed by S and square root calculation, N followed by S and fourth root calculation, N followed by autoscaling, N followed by the minimum-maximum linear transformation.

IS-normalization was considered as the reference approach since this preprocessing technique is routinely applied in the drug lab at the NICC to reduce instrumental variations, i.e. injection-volume variability and mass-spectrum variability. The other approaches were then compared to the reference with the aim to find the preprocessing technique that improves best the discriminating ability between linked and unlinked samples. The first eight numerical methods were based on published drug profiling techniques (26, 61). Autoscaling was used because this method creates equal importance for all variables. Linear transformation transforms the data with the aim to decrease the ratio of weight of very large peaks compared to small peaks. To evaluate the impact on distinction of the different numerical approaches, Pearson correlation coefficients were calculated, 95% and 99% confidence limits were determined on the preprocessed data set and the FN-FP rate was established. Cross-validation was carried out in two ways, i.e. leave-n-out cross-validation (LNO-CV) and leave-one-plantation-out cross-validation (LOPO-CV), to evaluate the predictive performance of the pretreatment resulting in the most improved discriminating method.

To evaluate the Pearson correlation coefficients, a “linked correlation” was defined as a correlation obtained from pairwise comparing chemical profiles originating from the same plantation, i.e. intra-location samples in the data set. “Unlinked correlations” were seen as computed correlation coefficients, between chemical profiles coming from different indoor cultivations, i.e. inter-location samples. Each intra-location correlation coefficient was determined, evaluated for possible outliers, and visualized in a colour map, representing the correlation coefficients matrices. For the third location, one sample could be distinguished (Fig. 6A), with lower  $r$ -values (blue colour) than the other samples which show high similarity (red-brown colour). Therefore, it was preferred to reject this outlying sample and to use only 9 samples to evaluate the variability within this plantation. However, it should be emphasized that a few intra-locations already show higher variability (Fig. 6B). Even though several

cannabis crops are cultivated under the same conditions, a higher variability among plants from one cultivation room is possible (65). Consequently, this naturally occurring variation also needs to be taken into account. The data set was thus reduced to 46 samples, resulting in a total of 1035 correlations, with 129 intra correlation coefficients and 906 inter correlation coefficients.

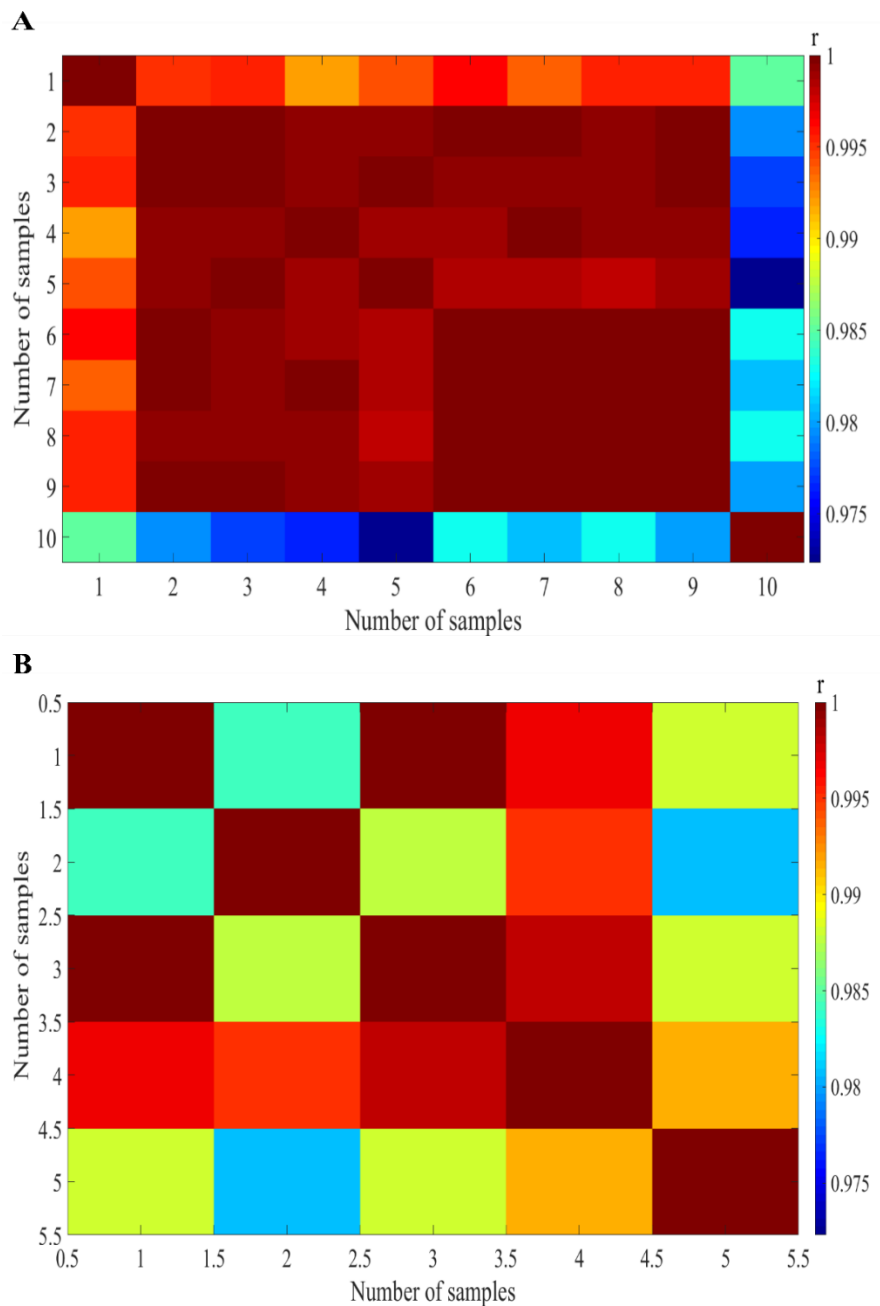


Fig 6. Correlation coefficients color map of (A) the third location showing an outlier (sample 10), and of (B) the second location showing heterogeneity between samples.

### 3.2.1 Discriminating ability evaluation

A critical value is derived from the intra-location samples using confidence interval lower limits (CIs). Two samples are considered linked when their correlation coefficient is above the lower limit. For a Gaussian distribution, 95% of the intra-location correlation coefficients is situated above the limit  $\bar{r} - 1.96 s_r$ , where  $\bar{r}$  is the average of the r-values of the linked samples and  $s_r$  the standard deviation of these linked correlation coefficients. 99% of the r-values is located above the limit  $\bar{r} - 2.576 s_r$ . The use of these limits theoretically results in constant FN percentages of 5% and 1%, respectively, as is shown in Fig. 7.

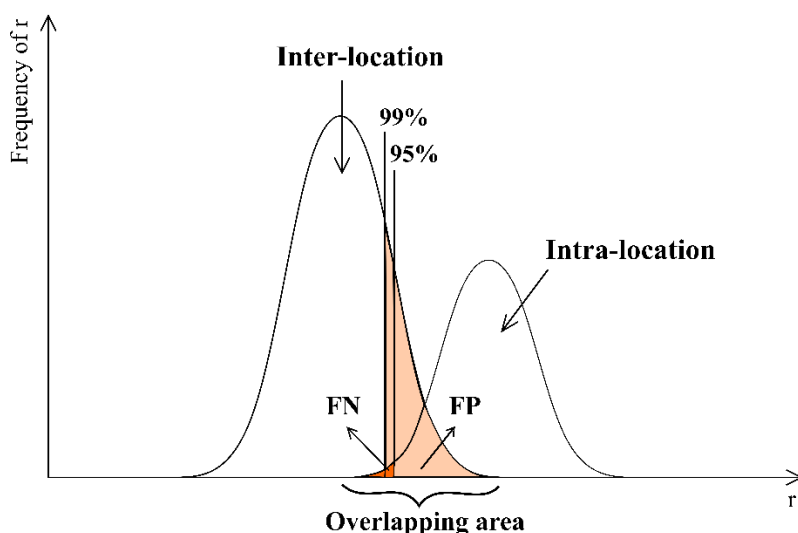


Fig 7. Theoretic inter-location and intra-location distributions of r, with indicated threshold values at the lower limits of the 95% and 99% confidence intervals of the intra-location distribution.

It is important to determine the FP rate, so a distinction between different plantations can be made. The goal of the chemical profiling data pretreatment is to minimize the overlap. Consequently, each pre-treatment applied will be evaluated by calculating the total % FNs and FPs. In table 4, the results are summarized for all transformations of the raw data with their respective total FN and FP rates. Initially, after IS normalisation, a threshold was determined of 0.986 for the 95% CI, showing 5% FNs and 21% FPs. Looking at the 99% CI, 3% FNs and 27% FPs were computed and a 0.983 threshold value was obtained. The aim is to find a method

with a FP rate below 10%. Thus the results that were found using the reference approach need to be further decreased. The majority of the preprocessed data sets demonstrated a decrease in % FPs compared to the reference pretreatment. However, the results show that the “N + row scaling to a constant total” pretreatment does not affect the separation between intra-location and inter-location distributions. N + log transformation generates an increase on the 95% level of FNs, comparing to the other preprocessing methods, with a threshold value of 0.993. This causes a higher intra-location dissimilarity. After applying N + fourth root normalisation, the FP rate was reduced to 11% and 14 % FPs, and to 3 % and 2 % FNs for both CIs. The computed thresholds are 0.994 and 0.992. This pre-treatment provides a better balance in the contribution of small and large peaks to the data analysis. The data were visualized with histograms showing the frequency of the Pearson correlation coefficients between intra- and inter-location samples for both the reference method and after N + fourth root normalization (Fig. 8). An overlap between the inter-location and intra-location distributions is seen and represents the capability of distinction between linked and unlinked samples. Comparison of Fig. 8a and 8b shows that the correlation-coefficient distributions are affected by the pretreatments and a noteworthy decrease in the overlapping region is observed in Fig. 8b.



Table 4. Total % FN-FP results, obtained at the two confidence levels, for the different data pretreatment methods applied.

Pre-treatment method	95% CL		99% CL	
	FN (%)	FP (%)	FN (%)	FP (%)
<b>IS-normalization (N)</b>	5	21	3	27
<b>N + square root</b>	4	14	3	18
<b>N + fourth root</b>	3	11	2	14
<b>N + Log transformation</b>	8	15	3	17
<b>N + row scaling constant total</b>	5	21	3	27
<b>N + standardisation (S)</b>	3	18	2	23
<b>N + S + square root</b>	2	15	2	19
<b>N + S + fourth root</b>	4	14	2	19
<b>N + auto-scaling</b>	5	13	3	16
<b>N + min-max transformation</b>	5	16	4	20

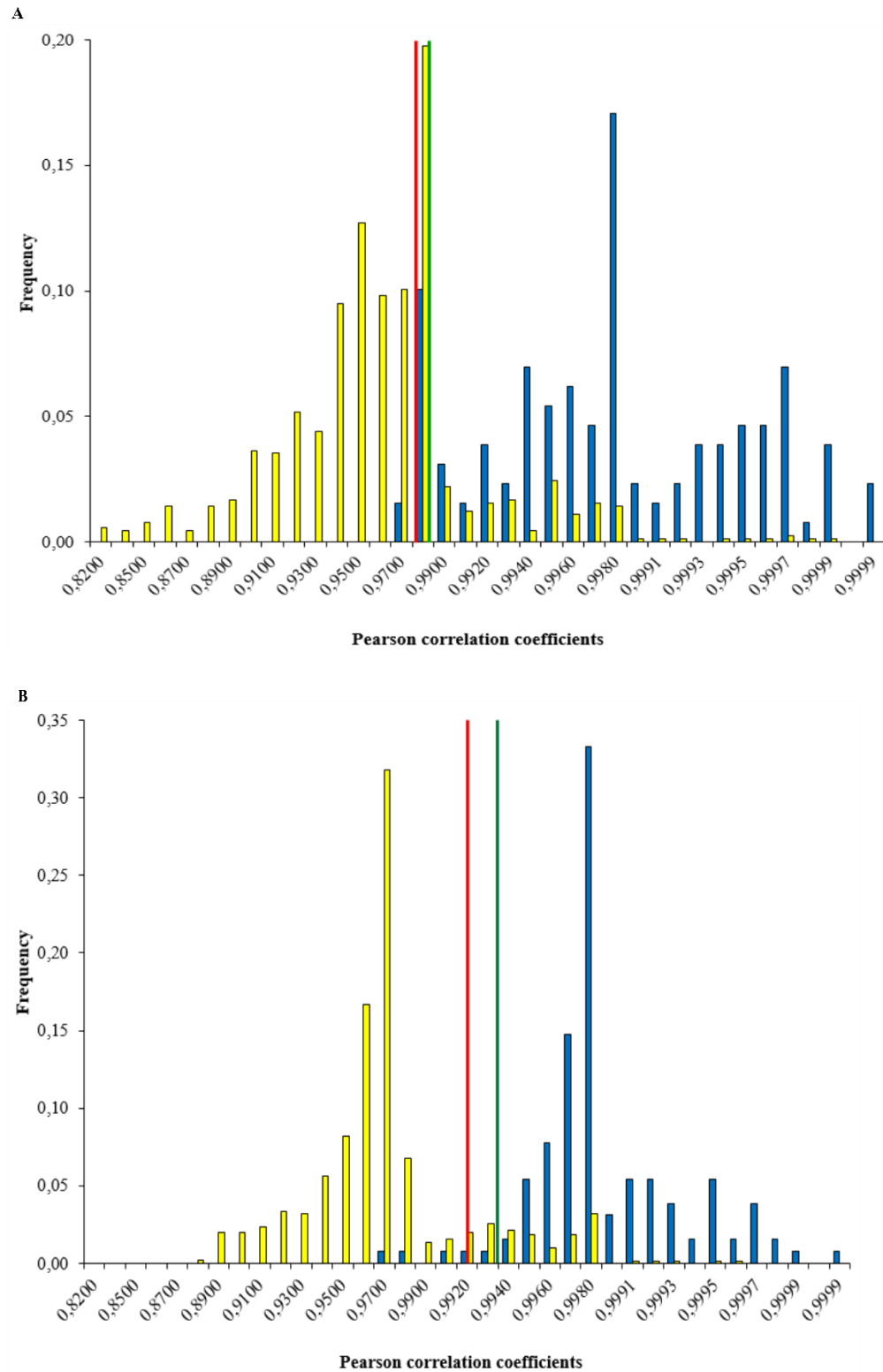


Fig 8. Histograms of the correlation coefficients after (a) IS-normalization and (b) N + fourth root normalization. The yellow bar chart corresponds to the coefficients of the inter-location samples, while the blue bar chart is derived from the intra-location correlation coefficients. Vertical lines: 95% CI (green) and 99% CI (red) limits.

An improved discrimination between the samples after data preprocessing can also be verified using ROC curves. The AUC of the N + fourth root normalization, i.e. 0.967, is higher than the AUC found with the reference method, i.e. 0.942 (Table 5). The respective ROC curves are shown in figure 9, where each point on the curve is a sensitivity- specificity combination that corresponds to a certain threshold. Here, the ROC curve representing the N + fourth root normalization increases closer towards the upper left corner, implying that the discriminating ability between intra- and inter-location samples is improved. Obtaining a better differentiation, when applying N + fourth root normalization is in agreement with what was reported earlier by Andersson et al. (27) and Morelato et al. (42).

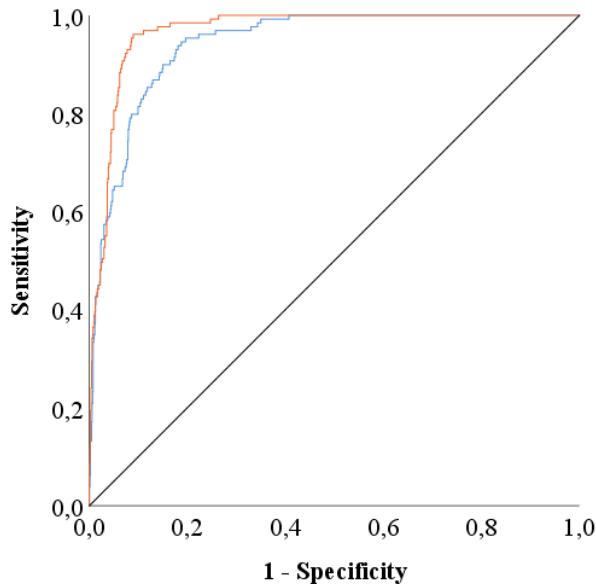


Fig 9. ROC curves demonstrating the discriminating ability of the two studied preprocessing methods, with the orange/red curve corresponding to the N + fourth root normalization and the blue the IS-normalization. The black line illustrates the reference line (AUC = 0.5).

Table 5. The respective AUCs for the reference method (IS-normalization) and after N + fourth root normalization.

Preprocessing	AUC	95 % confidence interval	
		Lower limit	Upper limit
IS-normalization	0,942	0,926	0,958
N + fourth root	0,967	0,957	0,977

### 3.3. Cross-validation

After evaluating all pre-treatments and concluding that “N + fourth root normalization” provides the most satisfying results, the predictive performance of the new approach will be evaluated and validated. Ideally, a calibration set is used to define the limits while an external test set provides a validation. Since the available data set consists of a low number of samples, cross-validation was applied instead to ensure that the defined limits are robust and accurate.

Two cross-validation approaches were performed:

- (1) LNO-CV: the data set consists of 1035 correlation coefficients. A training set and a test set are generated as disjoint subsets of the data set. First, the r-values were sorted in ascending order. The test set is obtained by leaving out 5% of the data in such way that each of the 20 test sets covers the entire r-range. This implies that each test group may contain intra- and inter-correlations for prediction and thus evaluation of both FNs and FPs can be done. The remaining r-values of the linked samples in the data set are used to define new CI limits. This procedure is repeated 20 times so that all coefficients are tested once and 20 limits were calculated.
- (2) LOPO-CV: With this approach, the data set is also subdivided in a set to define the limits and a validation set. The difference with the former cross-validation is that now the test set consists of one plantation. This was done to gain knowledge about particular cultivation sites. The remaining intra-plantation information of the 8 locations was used to define the 95% and 99% limits. This process is repeated 9 times, in order to eliminate

(3) all plantations once from the data set.

To test the accuracy in both approaches, results were validated by making predictions of the test set. The total % FN and FP error rates were derived from all test sets in each cross-validation approach, i.e. 20 for the first and 9 from the second. Table 6 gives an overview of the results from both setups. In both approaches, the total FP error rate was identical with 11% misclassified correlation coefficients, for the 95 % CI limit. For the 99% limit, both cross-validations showed approximately 14% FP. Concerning the % FN, a 6% error rate for the 95% threshold was obtained with LOPO-CV. This is twice as high as the value with LNO-CV but is still an acceptable value. An explanation for this observation is the higher variability occurring in certain plantations. In other words, limits were calculated without intra-values of a given plantation. It was already mentioned that samples within a few locations were more dissimilar compared to other intra-location samples causing lower intra-correlation coefficients. When determining both confidence interval limits, using these cultivation sites, lower limits were found. However, when these specific locations were used as test set, higher limits were achieved with the remaining training set, causing a higher % of false negatives for the test set. Interestingly, when comparing the predictive ability of the cross-validation approaches with the results obtained in section 3.3.1, where the total calibration data set was used to determine FNs and FPs, similar values were found. This indicated the representativeness of the defined limits.

Table 6. Total percentage false negatives/false positives (% misclassified correlation coefficients) of the two cross-validation approaches and the total calibration data set. Entire data set: 129 intra-r-values and 906 inter-r-values.

Cross-validation	% r-values misclassified			
	95 % CI limit		99 % CI limit	
	FN	FP	FN	FP
<b>Leave-n-out</b>	3	11	3	14
<b>Leave-one plantation-out</b>	6	11	4	13
<b>Entire calibration data set</b>	3	11	2	14

### 3.4. Conclusion

This paper focused on the effect of data pre-treatment methods and their ability to improve discrimination between linked and unlinked samples. Pearson correlation coefficients were used as pairwise similarity parameters for the chemical profiles. 95 % and 99% confidence limits were acquired based on intra-plantation information and used as threshold values. The internal standard normalization followed by fourth root transformation as preprocessing method was found to be most efficient to discriminate between different locations. The % FP was significantly lower (21 vs 11% for 95% CL) compared to the results obtained with the reference method (IS-normalization). Consequently, an improved approach with a reduced FN-FP overlap was generated. Cross-validation using 2 different methods, i.e. LNO-CV and LOPO-CV, showed similar FP results. compared to the calibration data set. The used data set was found representative to estimate the overall within-plantation variation. Future work will involve the use of a larger data set containing more intra-location samples to further investigate the intra-variability. Other approaches may also be studied with the goal to further decrease the % FP. The fingerprint profile as such, which represent all compounds in the cannabis matrix, could for instance, also be analyzed chemometrically. This entire-profile approach may then be compared to the above-applied peak-table method.

#### 4. References

1. De Backer, B., Debrus, B., Lebrun, P., Theunis, L., Dubois, N., Decock, L., et al. (2009) Innovative development and validation of an HPLC/DAD method for the qualitative and quantitative determination of major cannabinoids in cannabis plant material. *Journal of Chromatography B*, **877**, 4115–4124.
2. ElSohly, M.A. and Slade, D. (2005) Chemical constituents of marijuana: The complex mixture of natural cannabinoids. *Life Sciences*, **78**, 539–548.
3. Wang, Y.-H., Avula, B., ElSohly, M., Radwan, M., Wang, M., Wanas, A., et al. (2018) Quantitative Determination of  $\Delta^9$ -THC, CBG, CBD, Their Acid Precursors and Five Other Neutral Cannabinoids by UHPLC-UV-MS. *Planta Medica*, **84**, 260–266.
4. Hillig, K.W. and Mahlberg, P.G. (2004) A chemotaxonomic analysis of cannabinoid variation in *Cannabis* (Cannabaceae). *American Journal of Botany*, **91**, 966–975.
5. Stolker, A., Vanschoonhoven, J., Devries, A., Bobeldijkpastorova, I., Vaes, W. and Vandenberg, R. (2004) Determination of cannabinoids in cannabis products using liquid chromatography–ion trap mass spectrometry. *Journal of Chromatography A*, **1058**, 143–151.
6. United Nations Office On Drugs and Labor (2018) World Drug Report 2018 (Set of 5 booklets). United Nations, S.I. <https://www.unodc.org/wdr2018> (accessed on 30 September 2018).
7. European Monitoring Centre for Drugs and Drug Addiction (2018) European drug report 2018: trends and developments. <http://dx.publications.europa.eu/10.2810/800331>.
8. Vanhove, W., Surmont, T., Van Damme, P. and De Ruyver, B. (2012) Yield and turnover of illicit indoor cannabis (*Cannabis* spp.) plantations in Belgium. *Forensic Science International*, **220**, 265–270.
9. Cuypers, E., Vanhove, W., Gotink, J., Bonneure, A., Van Damme, P. and Tytgat, J. (2017) The use of pesticides in Belgian illicit indoor cannabis plantations. *Forensic Science International*, **277**, 59–65.
10. Hurley, J.M., West, J.B. and Ehleringer, J.R. (2010) Tracing retail cannabis in the United States: Geographic origin and cultivation patterns. *International Journal of Drug Policy*, **21**, 222–228.
11. Decorte, T. (2010) The case for small-scale domestic cannabis cultivation. *International Journal of Drug Policy*, **21**, 271–275.
12. McLaren, J., Swift, W., Dillon, P. and Allsop, S. (2008) Cannabis potency and contamination: a review of the literature. *Addiction*, **103**, 1100–1109.
13. Alvarez, A., Gamella, J.F. and Parra, I. (2016) Cannabis cultivation in Spain: A profile of plantations, growers and production systems. *International Journal of Drug Policy*, **37**, 70–81.

14. Toonen, M., Ribot, S. and Thissen, J. (2006) Yield of Illicit Indoor Cannabis Cultivation in The Netherlands. *Journal of Forensic Sciences*, **51**, 1050–1054.
15. Doran, G.S., Deans, R., De Filippis, C., Kostakis, C. and Howitt, J.A. (2017) Work place drug testing of police officers after THC exposure during large volume cannabis seizures. *Forensic Science International*, **275**, 224–233.
16. United Nations Office on Drugs and Crime (2009) Recommended methods for the identification and analysis of cannabis and cannabis products: manual for use by national drug testing laboratories. United Nations, New York. <http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=450863> (accessed on 30 September 2018).
17. ElSohly, M.A. (ed) (2007) Marijuana and the cannabinoids. Humana Press, Totowa, USA.
18. Collins, M., Huttunen, J., Evans, I. and Robertson, J. (2007) Illicit drug profiling: the Australian experience. *Australian Journal of Forensic Sciences*, **39**, 25–32.
19. Esseiva, P. and Margot, P. (2009) Drug Profiling. In Jamieson, A., Moenssens, A. (eds), Wiley Encyclopedia of Forensic Science. Wiley, Chichester, UK, p. fsa406. <http://doi.wiley.com/10.1002/9780470061589.fsa406> (accessed on 16 April 2019).
20. Esseiva, P., Ioset, S., Anglada, F., Gasté, L., Ribaux, O., Margot, P., et al. (2007) Forensic drug Intelligence: An important tool in law enforcement. *Forensic Science International*, **167**, 247–254.
21. Morelato, M., Beavis, A., Tahtouh, M., Ribaux, O., Kirkbride, P. and Roux, C. (2013) The use of forensic case data in intelligence-led policing: The example of drug profiling. *Forensic Science International*, **226**, 1–9.
22. Fraser, J.C. and Williams, R. (eds) (2009) Handbook of forensic science. Willan Publishers, Devon, UK.
23. Rannenbergh, K. (ed) (2009) The future of identity in the information society: challenges and opportunities. Springer, Berlin, Germany.
24. Inoue, H., Iwata, Y.T. and Kuwayama, K. (2008) Characterization and Profiling of Methamphetamine Seizures. *Journal of Health Science*, **54**, 615–622.
25. Esseiva, P., Dujourdy, L., Anglada, F., Taroni, F. and Margot, P. (2003) A methodology for illicit heroin seizures comparison in a drug intelligence perspective using large databases. *Forensic Science International*, **132**, 139–152.
26. Locicero, S., Esseiva, P., Hayoz, P., Dujourdy, L., Besacier, F. and Margot, P. (2008) Cocaine profiling for strategic intelligence, a cross-border project between France and Switzerland. Part II. Validation of the statistical methodology for the profiling of cocaine. *Forensic Science International*, **177**, 199–206.
27. Andersson, K., Lock, E., Jalava, K., Huizer, H., Jonson, S., Kaa, E., et al. (2007) Development of a harmonised method for the profiling of amphetamines VI. Evaluation of methods for comparison of amphetamine. *Forensic Science International*, **169**, 86–99.



28. Hewavitharana, A.K., Golding, G., Tempny, G., King, G. and Holling, N. (2005) Quantitative GC-MS Analysis of  $\Delta^9$ -Tetrahydrocannabinol in Fiber Hemp Varieties. *Journal of Analytical Toxicology*, **29**, 258–261.
29. Omar, J., Olivares, M., Amigo, J.M. and Etxebarria, N. (2014) Resolution of co-eluting compounds of *Cannabis Sativa* in comprehensive two-dimensional gas chromatography/mass spectrometry detection with Multivariate Curve Resolution-Alternating Least Squares. *Talanta*, **121**, 273–280.
30. Cardenia, V., Gallina Toschi, T., Scappini, S., Rubino, R.C. and Rodriguez-Estrada, M.T. (2018) Development and validation of a Fast gas chromatography/mass spectrometry method for the determination of cannabinoids in *Cannabis sativa* L. *Journal of Food and Drug Analysis*, **26**, 1283–1292.
31. Hazekamp, A. and Fishedick, J.T. (2012) Cannabis - from cultivar to chemovar: Towards a better definition of Cannabis potency. *Drug Testing and Analysis*, **4**, 660–667.
32. Gambaro, V., Dell'Acqua, L., Farè, F., Froidi, R., Saligari, E. and Tassoni, G. (2002) Determination of primary active constituents in Cannabis preparations by high-resolution gas chromatography/flame ionization detection and high-performance liquid chromatography/UV detection. *Analytica Chimica Acta*, **468**, 245–254.
33. Trofin, I.G., Vlad, C.C., Noja, V.V. and Dabija, G. (2012) Identification and characterization of special types of herbal cannabis. *U.P.B. Scientific Bulletin*, **74**, 119–130.
34. Vanhove, W., Van Damme, P. and Meert, N. (2011) Factors determining yield and quality of illicit indoor cannabis (*Cannabis* spp.) production. *Forensic Science International*, **212**, 158–163.
35. Stefanidou, M., Dona, A., Athanaselis, S., Papoutsis, I. and Koutselinis, A. (1998) The cannabinoid content of marihuana samples seized in Greece and its forensic application. *Forensic Science International*, **95**, 153–162.
36. Hazekamp, A., Tejkalová, K. and Papadimitriou, S. (2016) Cannabis: From Cultivar to Chemovar II—A Metabolomics Approach to Cannabis Classification. *Cannabis and Cannabinoid Research*, **1**, 202–215.
37. Potter, D.J. (2014) A review of the cultivation and processing of cannabis (*Cannabis sativa* L.) for production of prescription medicines in the UK: Cultivation and processing of cannabis for production of prescription medicines. *Drug Testing and Analysis*, **6**, 31–38.
38. Alaerts, G., Van Erps, J., Pieters, S., Dumarey, M., van Nederkassel, A.M., Goodarzi, M., et al. (2012) Similarity analyses of chromatographic fingerprints as tools for identification and quality control of green tea. *Journal of Chromatography B*, **910**, 61–70.
39. Fishedick, J.T., Hazekamp, A., Erkelens, T., Choi, Y.H. and Verpoorte, R. (2010) Metabolic fingerprinting of *Cannabis sativa* L., cannabinoids and terpenoids for chemotaxonomic and drug standardization purposes. *Phytochemistry*, **71**, 2058–2073.

40. Gröger, Th., Schäffer, M., Pütz, M., Ahrens, B., Drew, K., Eschner, M., et al. (2008) Application of two-dimensional gas chromatography combined with pixel-based chemometric processing for the chemical profiling of illicit drug samples. *Journal of Chromatography A*, **1200**, 8–16.
41. Liu, C., Hua, Z. and Meng, X. (2017) Profiling of illicit cocaine seized in China by ICP-MS analysis of inorganic elements. *Forensic Science International*, **276**, 77–84.
42. Morelato, M., Beavis, A., Tahtouh, M., Ribaux, O., Kirkbride, P. and Roux, C. (2014) The use of organic and inorganic impurities found in MDMA police seizures in a drug intelligence perspective. *Science & Justice*, **54**, 32–41.
43. Houck, M.M. (ed) (2015) Professional issues in forensic science. Elsevier/AP, Academic Press (an imprint of Elsevier), Oxford, UK ; San Diego, USA.
44. Cadola, L., Broséus, J. and Esseiva, P. (2013) Chemical profiling of different hashish seizures by gas chromatography–mass spectrometry and statistical methodology: A case report. *Forensic Science International*, **232**, e24–e27.
45. Brereton, R.G. (2009) Chemometrics for Pattern Recognition. Wiley, Chichester, UK. <http://doi.wiley.com/10.1002/9780470746462>.
46. Esseiva, P., Gaste, L., Alvarez, D. and Anglada, F. (2011) Illicit drug profiling, reflection on statistical comparisons. *Forensic Science International*, **207**, 27–34.
47. Houck, M.M. (ed) (2016) Materials analysis in forensic science. Academic Press/ Elsevier, Amsterdam ; Boston.
48. Li, W. and Liu, Z. (2011) A method of SVM with Normalization in Intrusion Detection. *Procedia Environmental Sciences*, **11**, 256–262.
49. Ratner, B. (2009) The correlation coefficient: Its values range between +1/–1, or do they? *Journal of Targeting, Measurement and Analysis for Marketing*, **17**, 139–142.
50. Puth, M.-T., Neuhausser, M. and Ruxton, G.D. (2014) Effective use of Pearson’s product–moment correlation coefficient. *Animal Behaviour*, **93**, 183–189.
51. Mukaka, M.M. (2012) Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal: The Journal of Medical Association of Malawi*, **24**, 69–71.
52. Zadora, G., Martyna, A., Ramos, D. and Aitken, C. (eds) (2014) Statistical analysis in forensic science: evidential value of multivariate physicochemical data. Wiley, Chichester, UK.
53. Zhou, H., Deng, Z., Xia, Y. and Fu, M. (2016) A new sampling method in particle filter based on Pearson correlation coefficient. *Neurocomputing*, **216**, 208–215.
54. Locicero, S., Hayoz, P., Esseiva, P., Dujourdy, L., Besacier, F. and Margot, P. (2007) Cocaine profiling for strategic intelligence purposes, a cross-border project between France and Switzerland. Part I. Optimisation and harmonisation of the profiling method. *Forensic Science International*, **167**, 220–228.

55. Tistarelli, M. and Champod, C. (eds) (2017) Handbook of Biometrics for Forensic Science. Springer International Publishing, Cham, Switzerland.
56. Siegel, J.A. (2007) Forensic science: the basics. CRC/Taylor & Francis, Boca Raton, USA.
57. Taroni, F. (ed) (2010) Data analysis in forensic science: a Bayesian decision perspective. Wiley, Chichester, UK.
58. Jackson, A.R.W. and Jackson, J.M. (2008) Forensic science. 2nd ed., Pearson Education, Harlow, UK.
59. Aitken, C.G.G. and Taroni, F. (2004) Statistics and the Evaluation of Evidence for Forensic Scientists: Aitken/Statistics and the Evaluation of Evidence for Forensic Scientists. 2nd ed., Wiley, Chichester, UK.
60. Broséus, J., Huhtala, S. and Esseiva, P. (2015) First systematic chemical profiling of cocaine police seizures in Finland in the framework of an intelligence-led approach. *Forensic Science International*, **251**, 87–94.
61. Marquis, R., Weyermann, C., Delaporte, C., Esseiva, P., Aalberg, L., Besacier, F., et al. (2008) Drug intelligence based on MDMA tablets data: 2. Physical characteristics profiling. *Forensic Science International*, **178**, 34–39.
62. Lasko, T.A., Bhagwat, J.G., Zou, K.H. and Ohno-Machado, L. (2005) The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, **38**, 404–415.
63. Likić, V.A. (2009) Extraction of pure components from overlapped signals in gas chromatography-mass spectrometry (GC-MS). *BioData Mining*, **2:6**. doi.org/10.1186/1756-0381-2-6.
64. Yang, J., Zhao, X., Lu, X., Lin, X. and Xu, G. A data preprocessing strategy for metabolomics to reduce the mask effect in data analysis. *Frontiers in Molecular Biosciences*, **2:4**. doi:10.3389/fmolb.2015.00004.
65. Sexton, M. and Ziskind, J. (2013) Sampling Cannabis for Analytical Purposes. 2013. [https://lcb.wa.gov/publications/Marijuana/BOTEC reports/1e-Sampling-Lots-Final.pdf](https://lcb.wa.gov/publications/Marijuana/BOTEC%20reports/1e-Sampling-Lots-Final.pdf) (accessed on 16 April 2019).