Journal of Chromatography A 1618 (2020) 460853

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Comprehensive characterisation of ylang-ylang essential oils according to distillation time, origin, and chemical composition using a multivariate approach applied to average mass spectra and segmented average mass spectral data

Leo Lebanov[a], Shiladitya Chatterjee[b], Laura Tedone[a], Sean C. Chapman[b], Matthew R. Linford[b], Brett Paull[a,*]

[a] Australian Centre for Research on Separation Science (ACROSS), School of Natural Sciences, University of Tasmania, Private Bag 75, Hobart 7001, Australia
[b] Department of Chemistry and Biochemistry, Brigham Young University, Provo, UT 84602, United States

ARTICLE INFO

Article history:
Received 26 August 2019
Revised 12 December 2019
Accepted 3 January 2020
Available online 9 January 2020

Keywords:
Average mass spectrum
Gas chromatography mass spectrometry
Ylang-ylang essential oil
Multivariate statistical analysis
Quality control

ABSTRACT

Analyses of the complex essential oil samples using gas chromatography hyphenated with mass spectrometry (GC–MS) generate large three-way data arrays. Processing such large data sets and extracting meaningful information in the metabolic studies of natural products requires application of multivariate statistical techniques (MSTs). From the GC–MS raw data several different input data sets for the MSTs can be created, including total chromatogram average mass spectra (TCAMS), segmented average mass spectra (SAMS) and chemical composition. Herein, we compared the performance of MSTs on average mass spectrum based data sets, TCAMS and SAMS, against chemical composition and attenuated total reflectance - Fourier transformation infrared (ATR-FTIR) spectroscopy in the evaluation of quality of ylang-ylang essential oils, based on their grade, geographical origin and chemical composition, using principal component analysis (PCA), partial least squares regression (PLS) and discriminatory analysis (PLS-DA). PCA based on TCAMS, SAMS and chemical composition showed clear trends amongst the samples based on increase in grade (distillation time). PLS-DA applied to TCAMS, SAMS and ATR-FTIR discriminated between all geographical origins. Predicted relative abundances of the 18 most important compounds, using PLS regression models on TCAMS, SAMS and ATR-FTIR, were successfully applied to ylang-ylang essential oil quality assessment based on comparison with the ISO 3063:2004 standard, where the SAMS data set showed superior performance, compared to other data sets.

1. Introduction

Essential oils (EOs) are mixtures of volatile and semi-volatile compounds obtained from natural sources by distillation, or by some suitable cold pressing method, such as in the case of citrus oils. Due to the complex and volatile nature of EO constituents, gas chromatography (GC) coupled with mass spectrometry (MS) is the most widely applied analytical technique for their characterisation [1,2]. However, the complexity of many EOs tests the resolution power of standard GC and often results in multiple co-eluting peak clusters [1]. Therefore, in recent years there has been an increase in the application of multidimensional gas chromatography (MDGC) in EO characterisation [1,3,4]. GC–MS and MDGC–MS generate large three- or multiway-way data arrays with high dimensionality (large number of variables and samples). Processing such large data sets, and extracting meaningful information in the metabolic studies of natural products requires application of multivariate statistical tools (MSTs) [5]. These techniques have been applied in EO metabolomic studies due to their ability to identify fine differences among the samples and classify them according to their species, chemotype (chemically distinctive entities of a population of morphologically indistinguishable individuals), or geographical origin [6,7].

Raw GC–MS data is a three-way array recorded as a function of three variables, time, *m/z* fragments and their intensities. There-

* Corresponding author.
E-mail addresses: leo.lebanov@utas.edu.au (L. Lebanov), laura.tedone@utas.edu.au (L. Tedone), mrlinford@chem.byu.edu (M.R. Linford), brett.paull@utas.edu.au (B. Paull).

https://doi.org/10.1016/j.chroma.2020.460853
0021-9673/

fore, data obtained in GC–MS analysis can be presented in several ways, through the dimensionality reduction process [8]. The most common is by summing all of the *m/z* fragments in a single mass scan and plotting them against the time of the mass scan, creating a total ion current chromatogram (TICC). Instrumental drifts in a GC–MS system can lead to small peak retention shifts between analyses. This means that before the application of an MST, it is necessary to perform peak alignment over all samples. In a TICC, alignment of peaks can be based solely on retention time or it can be performed using both retention time and mass spectra [9–12]. Another way of decreasing data dimensionality is by generating a chemical composition data set (peak list), as the MST input data set, that contains identified compounds and their relative abundances. Compounds are identified from the TICC, while the quantitative characterisation is performed on GC coupled with a flame ionization detector (FID). Finally, data obtained for all samples must be compiled into a single MST input data set. Due to manual tasks within this process, this whole approach can be very time consuming [13,14]. Several approaches have been proposed for the peak extraction and deconvolution from whole data sets, such as TUCKER3, PARAFAC, PARAFAC2, *etc*... While TUCKER3 and PARAFAC require trilinearity, that all the peaks are aligned, PARAFAC2 is able to handle small peak shifts between consecutive runs [15–17]. However, PARAFAC2 is not able to resolve all the peaks automatically in one step [18]. PARADISe is a software, based on PARAFAC2, for peak deconvolution which allows identification of the resolved pure MS spectra for each of the peaks. Before performing the deconvolution, narrow retention time intervals must be defined, which for complex samples can be very tedious. Also, after the deconvolution process, the number of individual components must be set for each of the retention time intervals. PARADISe requires at least 5 samples with independent chemical profile variations, in order to resolve the chemical profiles [17]. MARS is a fully automated peak deconvolution algorithm, which through coupled mass spectra and retention times, chosen from the reference chromatogram, extracts the chemical composition [18].

In contrast to the previous mentioned techniques for extracting the chemical composition profile as a MST input data set, there is an additional approach that averages the summed responses for each *m/z* fragment over the total number of scans, the whole time of the analysis, creating a total chromatogram average mass spectrum (TCAMS) data matrix [13,14,19]. Application of MSTs on TCAMS eliminates the necessity for peak identification and compiling a peak list, which makes obtaining the MST input data set much faster and easier [13]. It also allows for the creation of several data sets from the starting three-way data array recorded in an GC–MS analysis.

The downside of the application of TCAMS is that the actual chromatographic separation itself delivered by a GC column is lost. However, this does facilitate models using TCAMS being built upon data derived from the application of different instrument configurations and under different conditions of analysis, on different column setups, under different temperature gradients, etc. Another problem associated with the application of TCAMS can be interpretation of the models. Connecting/addressing *m/z* fragments to certain compounds in the chemical profile can be troublesome, as compounds in EOs belong to only a few chemical classes, and fragmentation of compounds which belong to the same chemical class have similar fragmentation patterns [20]. In order to exploit the advantages of TCAMS and preserve a certain level of chromatographic separation, a segmented average mass spectrum (SAMS) data reduction procedure is proposed in this work. SAMS is created by dividing the GC-MS three-way data array into time dependent sub-windows, where for each sub-window average mass spectrum (AMS) is calculated.

This study aimed to evaluate the principal component analysis (PCA), partial least squares regression (PLS) and discriminatory analysis (PLS-DA) performance of TCAMS and SAMS, against the chemical composition and attenuated total reflectance - Fourier transformation infrared spectroscopy (ATR-FTIR) results, as benchmark techniques [21,22], in the differentiation of Ylang-ylang (YY), *Cananga odorata* [Lam.] Hook. f. & Thomson forma genuina Annonaceae, EOs according to their grades and geographical origin. The YY EO obtained by fractional distillation of fresh, mature flowers is available in five different grades: Extra, I, II, III, and Complete, where with the increase in the distillation time, there is an increase in the less volatile fraction of sesquiterpenes, and a decrease of the highly volatile compounds. Complete grade is a mixture of all grades obtained in fractional distillation. The presence of the highly fragrant compounds, including linalool, *p*-cresyl methyl ether, methyl benzoate, benzyl acetate, and geranyl acetate, make the Extra grade extremely popular in the perfume industry, while grade I with lower amounts of volatile compounds is more applied within the cosmetic industry. Madagascar and the Comoros Islands are the world's largest producers of this EO, where the value of the total production of YY EO in the Union of Comoros was 7.9 M USD in 2008, making it the largest contributor to EO production in that country [23,24]. Herein, TCAMS, SAMS and ATR-FTIR were compared in the prediction of the chemical composition of the 18 most important compounds in the assessment of the quality of YY EOs according to the ISO 3063:2004 "Oil of ylang-ylang [*Cananga odorata* (Lam.) Hook. f Thomson forma *genuine*]" standard.

2. Experimental

2.1. Samples

YY EOs of varying geographic origin and grades (see Table 1), including Extra, I, II, III, and Complete grades from Madagascar, Comoros, Nepal, and/or France were obtained from Plant Therapy Inc. (Twin Falls, ID, USA). YY Complete EOs from Australia, Congo, and Indonesia, and YY Extra EOs from Comoros were obtained from Ahimsa Oils (Mitchelton, QLD, Australia). Two samples of YY Complete from Madagascar and YY I from Comoros were obtained from Albert Vieille SAS (France). Each sample was prepared in the same manner, by diluting 50 µL of the pure EO in 950 µL of dichloromethane (DCM). HPLC-grade DCM was purchased from Thermo Fisher Scientific (Waltham, MA, USA). A standard n-alkane mixture C_7-C_{30} was purchased from Sigma-Aldrich (St. Louis, MO, USA).

2.2. GC–qMS analysis

Samples were analysed with a GCMS-QP2010 Plus instrument, Shimadzu (Kyoto, Japan) with a BPX-5MS capillary column (33.0 m × 0.25 mm ID × 0.25 µm) purchased from SGE, Australia. The sample, 1 µL, was injected at 250 °C in split mode with a 50:1 ratio. The oven temperature program was set as follows: 50 °C to 250 °C at 4.30 °C/min, and from 250 °C to 280 °C (5 min hold) at 15.0 °C/min. Ultra-high purity helium (BOC, UK) was used as the carrier gas and the flow rate was constant throughout the run (1.54 mL/min, linear velocity of 42.9 cm/s). The ion source was set at 220 °C with the transfer line at 250 °C. Data were collected with an acquisition rate of 0.2 s with a scan speed of 2000 over a mass range of 40–400 Da.

Identification of the EO components was performed using similarity match (≥90%) between mass spectrum of each compound with the standard spectrum, together with comparing their retention indices (tolerance of ±10) with the standard value. The similarity search was performed using GCMS solution software (version 4.30, Shimadzu, Kyoto, Japan) within the standard libraries

Table 1
List of samples used in this work.

Code	Grade	Origin	Obtained from	Code	Grade	Origin	Obtained from
YExtraMad01	Extra	Madagascar	Plant Therapy	YIIIMad01	III	Madagascar	Plant Therapy
YExtraMad02	Extra	Madagascar	Plant Therapy	YIIIMad02	III	Madagascar	Plant Therapy
YExtraMad03	Extra	Madagascar	Plant Therapy	YIIIMad03	III	Madagascar	Plant Therapy
YExtraNep04	Extra	Nepal	Plant Therapy	YIIIMad04	III	Madagascar	Plant Therapy
YExtraMad05	Extra	Madagascar	Plant Therapy	YIIIMad05	III	Madagascar	Plant Therapy
YExtraCom06	Extra	Comoros	Plant Therapy	YIIIMad06	III	Madagascar	Plant Therapy
YExtraMad07	Extra	Madagascar	Plant Therapy	YIIIMad07	III	Madagascar	Plant Therapy
YExtraCom08	Extra	Comoros	Ahimsa Oils	YIIICom08	III	Comoros	Plant Therapy
YExtraCom09	Extra	Comoros	Ahimsa Oils	YComplFra01	Complete	France	Plant Therapy
YExtraCom10	Extra	Comoros	Ahimsa Oils	YComplMad02	Complete	Madagascar	Plant Therapy
YExtraCom11	Extra	Comoros	Ahimsa Oils	YComplMad03	Complete	Madagascar	Plant Therapy
YIFra01	I	France	Plant Therapy	YComplMad04	Complete	Madagascar	Plant Therapy
YIMad02	I	Madagascar	Plant Therapy	YComplMad05	Complete	Madagascar	Plant Therapy
YIMad03	I	Madagascar	Plant Therapy	YComplMad06	Complete	Madagascar	Plant Therapy
YIMad04	I	Madagascar	Plant Therapy	YComplMad07	Complete	Madagascar	Plant Therapy
YICom05	I	Comoros	Plant Therapy	YComplAus08	Complete	Australia	Ahimsa Oils
YIMad06	I	Madagascar	Plant Therapy	YComplAus09	Complete	Australia	Ahimsa Oils
YIMad07	I	Madagascar	Plant Therapy	YComplAus10	Complete	Australia	Ahimsa Oils
YIMad08	I	Madagascar	Plant Therapy	YComplAus11	Complete	Australia	Ahimsa Oils
YICom09	I	Comoros	Albert Vieille	YComplCng12	Complete	Congo	Ahimsa Oils
YIIFra01	II	France	Plant Therapy	YComplCng13	Complete	Congo	Ahimsa Oils
YIIMad02	II	Madagascar	Plant Therapy	YComplCng14	Complete	Congo	Ahimsa Oils
YIIMad03	II	Madagascar	Plant Therapy	YComplCng15	Complete	Congo	Ahimsa Oils
YIICom04	II	Comoros	Plant Therapy	YComplMad16	Complete	Madagascar	Albert Vieille
YIIMad05	II	Madagascar	Plant Therapy	YComplInd17		Indonesia	Ahimsa Oils
YIIMad06	II	Madagascar	Plant Therapy	YComplInd18		Indonesia	Ahimsa Oils
				YComplInd19		Indonesia	Ahimsa Oils
				YComplInd20		Indonesia	Ahimsa

of the National Institute of Standards and Technology (NIST, 2005), Wiley (Wiley Registry™ of Mass Spectral Data, 11th Edition, 2016), and Flavours and Fragrances of natural and synthetic compounds – Mass spectral database (FFNSC 2.0, 2007). Retention indices (RIs) were calculated based on the linear temperature programme retention index (LTPRI) procedure by using a standard linear alkane mixture of C_7-C_{30} established by Van den Dool and Kratz for the temperature programmed separation [25].

2.3. GC-–FID

Samples were analysed with an Agilent 7890A instrument (Palo Alto, CA, USA), employing a BP5MS column, 30 m × 0.25 mm × 0.25 µm, purchased from SGE, Australia. The sample, 1 µL, was injected at 250 °C in split mode (1:20). The oven temperature program was set as follows: 50 °C to 250 °C at 4.3 °C/min, and from 250 °C to 280 °C (5 min hold) at 15 °C/min. Ultra-high purity H_2 was used as the carrier gas and the flow rate was held constant throughout the run (1.54 mL/min, linear velocity of 42.7 cm/s). The FID detector was set at 280 °C, with a hydrogen flow of 40 mL/min and an air flow of 450 mL/min. Data processing was performed using the GC ChemStation B.0403 (16) software.

2.4. ATR-FTIR spectroscopy

ATR-FTIR spectra were measured using a Bruker Vertex 70 instrument (Germany) at the Central Science Laboratory (CSL) at the University of Tasmania. Spectra were collected by placing 0.8 µL of oil on a PIKE (USA) MIRacle single reflection ZnSe ATR plate, using 32 scans between 600 and 4000 cm^{-1} at a spectral resolution of 4 cm^{-1}. The spectra were corrected using the atmospheric correction function within the OPUS software (Bruker, Germany).

Multivariate analysis was performed by importing the spectral data into Unscrambler (v 10.1, CAMO, Norway).

2.5. Data processing

TCAMS of each sample were taken from 2.5–40.0 min while subtracting the baseline from 39.0–40.0 min. SAMS were taken at 5 different regions: 2.5–10.0 min, 10.0–15.0, 15.0–20.0, 20.0–25.0, and 25.0–40.0 min, with baseline subtraction for each segment. The number of segments was determined to preserve the separation achieved and correct for any possible peak shifts. Therefore, it was determined that 5 was the optimal number for segments, as this best captured the chemical diversity of essential oils, where small volatile molecules and monoterpenes elute in the first segment, monoterpenoids in the second, sesquiterpenes in the fourth, and sesquiterpenoids in the fifth. Also, 5 segments allowed correction for small peak shifts which arise between consecutive analyses. Several compounds were used as markers for the segments. Segment 1 included p-cresyl methyl ester and limonene, co-eluting with 1,8-cineole, which were taken as markers for the end of Segment 1. Estragole was applied as a segment 2 end marker. Geranyl acetate was used as a starting compound for Segment 4, ending with the sesquiterpene zonarene. Segment 5 included all sesquiterpenoids, eluting after zonarene.

Raw data were processed by metaboanalyst.ca [26], which is online software based on XCMS code written in the R software package, by eliminating the *m/z* fragments that are constant through all the samples (noise). Next, the data were pre-processed by mean centering and normalisation (1-norm). Normalisation was performed to correct for inconsistencies in the amounts of samples reaching the detector, while keeping the structure of the data. PCA was performed with metaboanalyst.ca, while PLS was performed using Sirius (v. 9.0, PRS, Norway) and PLS-DA using the *mdatools* in R software [27]. The ATR-FTIR data were pre-processed with the Uncrambler software, which included a multiple scattering correction and Savitzky-Golay first derivative using 7 points.

2.5.1. Theory of TCAMS, SAMS and TICC

If m is the total number of mass scans (S_i) recorded during a single GC–MC run in time, t, with Δt the time necessary to record one mass scan, then $m = {}^{t}/_{\Delta t}$. Here, n is the number of m/z fragments (40–400) and y_{ij} is the response for each m/z fragment in each scan, where i = 1,2,...,m and j = 1,2,...,n.

The mass scan S_i (i = 1,2,...,m) could be defined as a set of m/z fragments with their corresponding responses from the MS detector:

$$S_i = \left\{\left[y_{i1},\ \left({}^{m}/_{z}\right)_1\right], \left[y_{i2},\ \left({}^{m}/_{z}\right)_2\right], \ldots, \left[y_{in},\ \left({}^{m}/_{z}\right)_n\right]\right\}. \quad (1)$$

The TICC (see Fig. 1) can be defined as

$$\begin{aligned} TICC = \{&[(y_{11} + y_{12} + \ldots + y_{1n}), \Delta t], \\ &[(y_{21} + y_{22} + \ldots + y_{2n}), 2\Delta t], \ldots, \\ &[(y_{m1} + y_{m2} + \ldots + y_{mn}), m\Delta t]\}. \end{aligned} \quad (2)$$

A TCAMS (see Fig. 1) can be defined as an average response of each m/z fragment during the whole GC analysis

$$\begin{aligned} TCAMS &= \left\{\left[\bar{y}_1, \left({}^{m}/_{z}\right)_1\right], \left[\bar{y}_2, \left({}^{m}/_{z}\right)_2\right], \ldots, \left[\bar{y}_n, \left({}^{m}/_{z}\right)_n\right]\right\} \\ &= \left\{\left[\frac{(y_{11,} + y_{21} + \ldots + y_{m1})}{m}, \left({}^{m}/_{z}\right)_1\right], \right. \\ &\quad \left[\frac{(y_{12,} + y_{22} + \ldots + y_{m2})}{m}, \left({}^{m}/_{z}\right)_2\right], \ldots, \\ &\quad \left.\left[\frac{(y_{1n,} + y_{2n} + \ldots + y_{mn})}{m}, \left({}^{m}/_{z}\right)_n\right]\right\}. \end{aligned} \quad (3)$$

Here, $\bar{y}_j$ is the average response of the MS for a single $(m/z)_j$ ion current over the whole scanning period t, where all recorded scans were taken into account.

A SAMS can be defined as an average response of each m/z fragment in the defined time window (segment) of the GC analysis, where all segments are compiled into a single data set (see Fig. 1). In the case of SAMS, for k segments, m_k represents the number of scans in the defined kth segment with time t, $m_k = {}^{t_k}/_{\Delta t}$. A SAMS can be defined as an average response of each m/z fragment in the defined segment time window (t_p) of the GC–MS analysis over all the segments (k, total number of segments). The *AMS* of each segment can be defined in the same manner as the TCAMS, where the number of scans in each segment is defined as $m_p = {}^{t_p}/_{\Delta t}$, where p = 1,2...k, and y_{ijp} is the response for each m/z in each scan and in each segment.

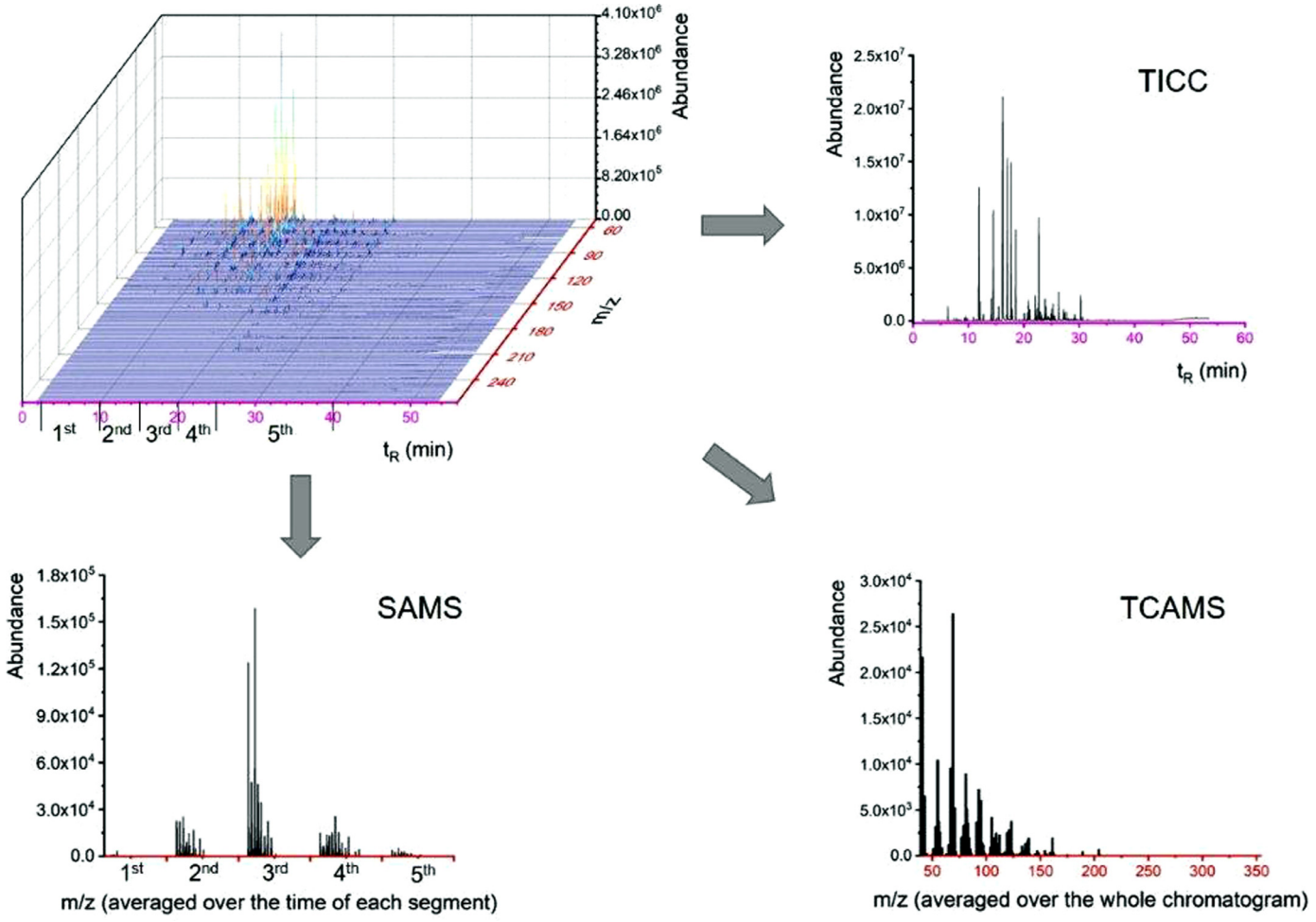

Fig. 1. Reduction of dimensionality of a three-way GC-MS data to total ion current chromatogram (TICC), total chromatogram average mass spectrum (TCAMS) and segment average mass spectrum (SAMS) data sets. GC-MS data are recorded in three dimensions: time, m/z fragments and their abundance. For the TICC, all m/z fragments in the same scan are summed; the TCAMS is created by averaging the response of each m/z fragment over a chromatographic run; the SAMS is created by averaging the response for each m/z fragment over the time of a chromatogram segment. From 5 segments, 5 average mass spectra data sets were created, which are compiled into a single data array. Reproduced with permisson from [20].

$$\mathrm{SAMS} = \{[\mathrm{AMS}_1, (1)], [\mathrm{AMS}_2, (2)], \ldots, [\mathrm{AMS}_p, (p)]\} = \left\{ \begin{array}{l} \left[\left[\frac{(y_{111} + y_{211} + \ldots + y_{m11})}{m_1}, (m/z)_1\right], \left[\frac{(y_{121} + y_{221} + \ldots + y_{m21})}{m_1}, (m/z)_2\right], \ldots, \left[\frac{(y_{1n1} + y_{2n1} + \ldots + y_{mn1})}{m_1}, (m/z)_n\right]\right], \ldots, \\ \left[\left[\frac{(y_{11p} + y_{21p} + \ldots + y_{m1p})}{m_p}, (m/z)_1\right] \left[\frac{(y_{12p} + y_{22p} + \ldots + y_{m2p})}{m_p}, (m/z)_2\right], \ldots, \left[\frac{(y_{1np} + y_{2np} + \ldots + y_{mnp})}{m_p}, (m/z)_n\right]\right] \end{array} \right\} \quad (4)$$

As can be observed, the TICC is a time dependent function, i.e., TICC(t) is the summation of the abundances of all the recorded m/z fragment intensities in a single scan. Thus, the TICC response is proportional to the sum of responses of all the m/z fragments in a single scan, which is proportional to the abundance of compounds producing those fragments. On the other hand, the TCAMS is the average response of the mass selective detector during the whole GC–MS run (Δt - $m\Delta t$) as a function of m/z (single ion current). The average response of each mass fragment is proportional to the abundance of the specific m/z fragments obtained, which are again proportional to the abundance of the compounds producing these ions. Therefore, the TICC, TCAMS, and SAMS responses depend on the chemical composition of the EO and the amount of injected analyte that can be used with the MST in metabolomic studies and the prediction of the relative abundances of the main compounds.

2.5.2. Principal component analysis (PCA) and partial least squares (PLS)

PCA is a statistical, unsupervised pattern recognition technique that is widely used for dimensionality reduction, where trends and patterns among the sample set can be more easily identified in a space of lower dimension, e.g., two or three-dimensional space, which is in detailed explained elsewhere [28–30]. PCA has been widely used for dimensionality reduction, especially in the case of complex instrumental responses [31].

PLS is a two-block predictive chemometric technique that attempts to establish a *cause and effect* relationship using a set of predictor variables (**X**) and response variables (**Y**) [32,33]. PLS reduces data dimensionality by calculating pseudo variables (latent variables) that attempt to maximise the covariance between the response and matrix of the predictor variables [32,33]. Additional explanations can be found elsewhere [33].

In the case herein, instrumental raw data obtained by GC–MS and ATR-FTIR were used as predictor variables, while grade, geographical origin and chemical composition were the dependent (response) variables. PLS regression analysis was performed in case of grade and chemical composition determination, while PLS-DA was used for geographical origin determination. In this case grade regression was used, with dummy variables 1 for grade Extra, 2 for grade I, 3 for grade II and 4 for grade III. In this way, it was possible to evaluate the fragrant character of each of the samples and compare with the other samples, on a scale from 1–4. This approach has advantage for the evaluation of borderline samples, whose fragrant characters are similar and can be equally used either in the perfume or cosmetic industries.

Performance of PLS and PLS-DA models was assessed by performing leave-one-out cross-validation (LOO CV), which leaves one sample out of the training set in each run, creates a model of the training set, predicts the value of the left-out sample, and calculates the root mean square error of cross-validation (RMSECV). The goodness of the PLS model created from the training set can be evaluated through the root mean square error of calibration (RMSEC) value:

$$\mathrm{RMSEC} = \sqrt{\frac{\Sigma_{i=1}^{N}\left(C_i - C_i'\right)^2}{N - 1 - p}}. \quad (5)$$

where C_i are the known values, C_i' are the calculated (predicted) values, N is the number of samples, and p is the number of PCs in the regression analysis optimised by cross validation. The value -1 in the denominator accounts for a centred model. RMSECV estimates the prediction performance of cross validation applying the equation:

$$\mathrm{RMSECV} = \sqrt{\frac{\Sigma_{i=1}^{M}\left(C_i - C_i'\right)^2}{M}}. \quad (6)$$

where C_i is the known value, C_i' is the value calculated (predicted) by the calibration equation, and M is the number of samples in the prediction set.

Another parameter for the evaluation of the model is the relative error of prediction (REP). The REP shows the predictive ability of the model and can be calculated as:

$$\mathrm{REP} = \frac{\mathrm{RMSECV}}{\bar{y}} \times 100. \quad (7)$$

where $\bar{y}$ is the mean of the predicted values. PLS-DA models for geographical origin discrimination were assessed using a specificity and sensitivity confusion matrix [34,35].

Another way of evaluating the goodness in the PLS model is by performing a permutation test, explained elsewhere [35,36]. In the case of our PLS models, the permutation was repeated 2000 times with $P \leq 0.05$, showing that our sets of samples are non-permuted (results not shown here). The validation of the model can be expressed as the correlation coefficient (R^2), slope and intercept of the created calibration curve on the training sample set.

Evaluation of the quality of the application of the predicted values for the quality assessment of YY EOs was found by determining the percentage of misclassification. Percentage of misclassification (MC) is a sum of all false positives and negatives divided by the total sum of all the compounds (*TNC*) determined in all the samples and multiplied by 100%.

$$\mathrm{MC} = \frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{TNC}} \times 100\%. \quad (8)$$

A false positive (*FP*) is when the relative abundance of the compound is outside of the ISO range, while the predicted value utilising the PLS model falls inside of the ISO range. A false negative (*FN*) can be determined when the relative abundance of the compound is inside of the range, but the predicted value is outside of the ISO range for that compound [36]. The quality of the application of the PLS prediction model was evaluated by taking into account both the MC and REP of each model.

3. Results and discussion

3.1. Two-way data arrays obtained by GC–MS

As already stated, the GC–qMS analysis results in an array of three-way data. In order to apply MSTs, these data matrices were processed in three different ways for dimensionality reduction, creating chemical composition, TCAMS, and SAMS data sets. The chemical composition data set was constructed of 83 main compounds through all the samples. In this way, a full chemical profile was obtained for each sample, and by performing the exploratory

and prediction analysis, the main compounds responsible for the clustering and which influence the prediction can be easily identified.

TCAMS and SAMS are directly dependent on the chemical composition of the whole sample, through *m/z* fragment abundances. Thus, samples that belong to the same species show certain homogeneity in the abundances of *m/z* fragments in the TCAMSs between different YY grades (see Fig. S1.). At the same time, differences in the chemical profiles between the samples lead to variability of the YY samples, which are translated into differences in *m/z* abundances.

TCAMS can also be obtained by performing direct injection mass spectrometry (DI-MS) analysis. However, application of a separation step before acquiring TCAMS has an additional benefit compared to application solely of the DI-MS system. Application of a long bore column decreases the complexity and mass of the sample reaching the MS detector, compared to direct MS injection, and decreases the possibility of saturation of certain channels, which can lead to the introduction of additional variation in the data sets that does not come from the variation between the samples [14]. Another advantage of using GC–MS is the possibility of using SAMS data, which retain elements of a chromatographic separation, by dividing the raw three-way data arrays into chromatographic segments (sub-windows), where for each segment the AMS is generated. AMSs from all segments are compiled into a single data matrix called a segment average mass spectrum (SAMS). Keeping certain elements of the separation, there is a significant increase in the number of data points (variables), compared to the TCAMS, where their abundance in a single segment depends on the compounds eluting within that segment. This allows the model to evaluate the importance of each *m/z* fragment from all the segments, and possibly increase predictability and accuracy of the model. In this way, interpretation of the models becomes much easier, which allows identification of important fragments in certain segments and connects them with the chemistry of the eluting compounds therein. SAMS does not require strict peak alignment. However, the same compounds and peaks must be present in the same segment, which is achieved through marker classes and compounds [20]. Compared to the chemical composition created by PARAFAC2 and PARADISe, with SAMS the retention intervals are larger and there is no necessity for defining the number of individual compounds in each interval (segment). Also, there is no minimal number of samples necessary to create a working data set, or necessity for a reference sample, as required in the MARS algorithm. Utilising the AMS approach, in both TCAMS and SAMS, chemical composition is translated into a series of *m/z* fragments, the same for each sample, which decreases the number of missing values.

3.2. Determination of the ylang-ylang EO grade

3.2.1. Exploratory analysis

Before constructing PLS prediction models, unsupervised exploratory analysis was performed [37]. Performing PCA analysis on all the samples from four different distillation grades using two PCs, it was observed that two samples (YExtraMad02 and YIIMad02) clustered with the grade I, Fig. S2. Evaluating their chemical profiles, it was clear that YIIMad02 had lower amounts of germacrene D, compared to the ISO 3063:2004 standard [38] for the grade II samples. YExtraMad02 clusters with the grade I samples due to the lower amount of *p*-cresyl methyl ether and methyl benzoate and higher abundance of germacrene D, which is also below the ISO 3063:2004 standard [38] reference values for the Extra grade. Therefore, due to the differences in the chemical profile from the other grade-members, these samples were excluded from all PCA and PLS models.

In order to construct an adequate PCA model, scree plots were examined to determine the optimum number of PCs (Figs. S3 A-D) [39]. A scree plot displays the eigenvalues (explained variance) for each PC in descending order versus the number of components. The optimal number of PCs can be obtained (estimated) by taking the last PC at which point little additional variance is captured by the PCs, which is often indicated by a so-called 'knee' in the plot.

Performing PCA on the TCAMS data (Fig. 2A), it was observed that the grade III cluster was completely separated from the rest of the samples. Along PC 1, which explains 55.9% of the variance in the data set, the grades Extra and I samples are closely clustered. The higher PCs were unable to contribute to an increase in the separation between the grade Extra and I samples (data not presented here), even though the scree plot shows a greater amount of variance explained by them, see Fig. S3A. Therefore, an optimal PCA model was achieved with 2 PCs. SAMS, compared to TCAMS, retains aspects of the chromatographic separation, which is reflected in the increased number of variables. In the case of SAMS, a PCA model with one PC was suggested by the scree plot (Fig. S3B). As can be observed in Fig. 2B, PCA with only the first PC (PC 1) can discriminate between YY grades based on the SAMS data set. In the PCA utilised on the chemical composition (Fig. 2C), the two clusters corresponding to the grade II and III samples were better resolved as compared to within the TCAMS and SAMS data sets. The Extra and I samples are less overlapped along PC 1 and PC 3 (Fig. 2D). For the ATR-FTIR data set, the scree plot, Fig. S3D, suggested just PC 1 in the model, which explained 97% of the total variance, even though it is not possible to resolve any grade, even within a two-dimensional PCA plot, see Fig. 2E.

The longer distillation, in the case of grade III, creates samples with distinctive chemistry, which results in well resolved clusters during PCA analysis of the TCAMS, SAMS, and chemical composition (see Fig. 2A-D) data sets. As is evident from the raw TICCs of the representative samples from each of the 4 grades (Figs. S4A-D), the Extra and I grade TICCs appear almost identical, which was reflected in the closeness of these two grades in the PCA analyses (Fig. S4A, B). These samples were provided by a commercial vendor, and there was a lack of control over the flower picking and distillation process, which can be attributed to the similarity between the TICCs for the Extra and I grades.

The loadings for the PCA model of the TCAMS data (Fig. 3A) show that the primary *m/z* fragments responsible for separation of the higher grades are 119 and 161 (sesquiterpenes), while the 43, 71, 77, 108, and 122 fragments are connected with the lower grades. Fragments 71, 108, and 122 are characteristic of linalool, benzyl acetate, and *p*-cresyl methyl ether, respectively. Fragment 43 is associated with linalool, geranyl acetate, and benzyl acetate, while fragment 77 is representative of methyl benzoate, benzyl acetate, benzyl benzoate, and *p*-cresyl methyl ether. From the loadings plot of the PCA analysis of the SAMS data (Fig. 3B), it was observed that in the higher grades there was an increase in *m/z* fragments 41, 91, 93, 105, etc., from the fourth segment, while an increase in the *m/z* fragments from the second segment, 43, 71, 77, 91, 93, 105, and 108, is indicative of the lower YY grade. Retention indices 1484 (germacrene D) and 1503 ((*E,E*)-α-farnesene) were the main compounds in the higher grades, while 1016 (*p*-cresyl methyl ether), 1089 (methyl benzoate), 1096 (linalool), 1155 (benzyl acetate), and 1373 (geranyl acetate) were more abundant in the lower grades in the loadings for chemical composition (Fig. 3C). Moreover, β-caryophyllene (RI 1423) is connected with the variations within the YY grades, Fig. 3.

3.2.2. PLS Prediction model of ylang-ylang EO grade

The PCA results described in the previous section reveals that distillation time has a great impact on YY EO chemical composition and that it can be used for constructing a PLS regression model. At

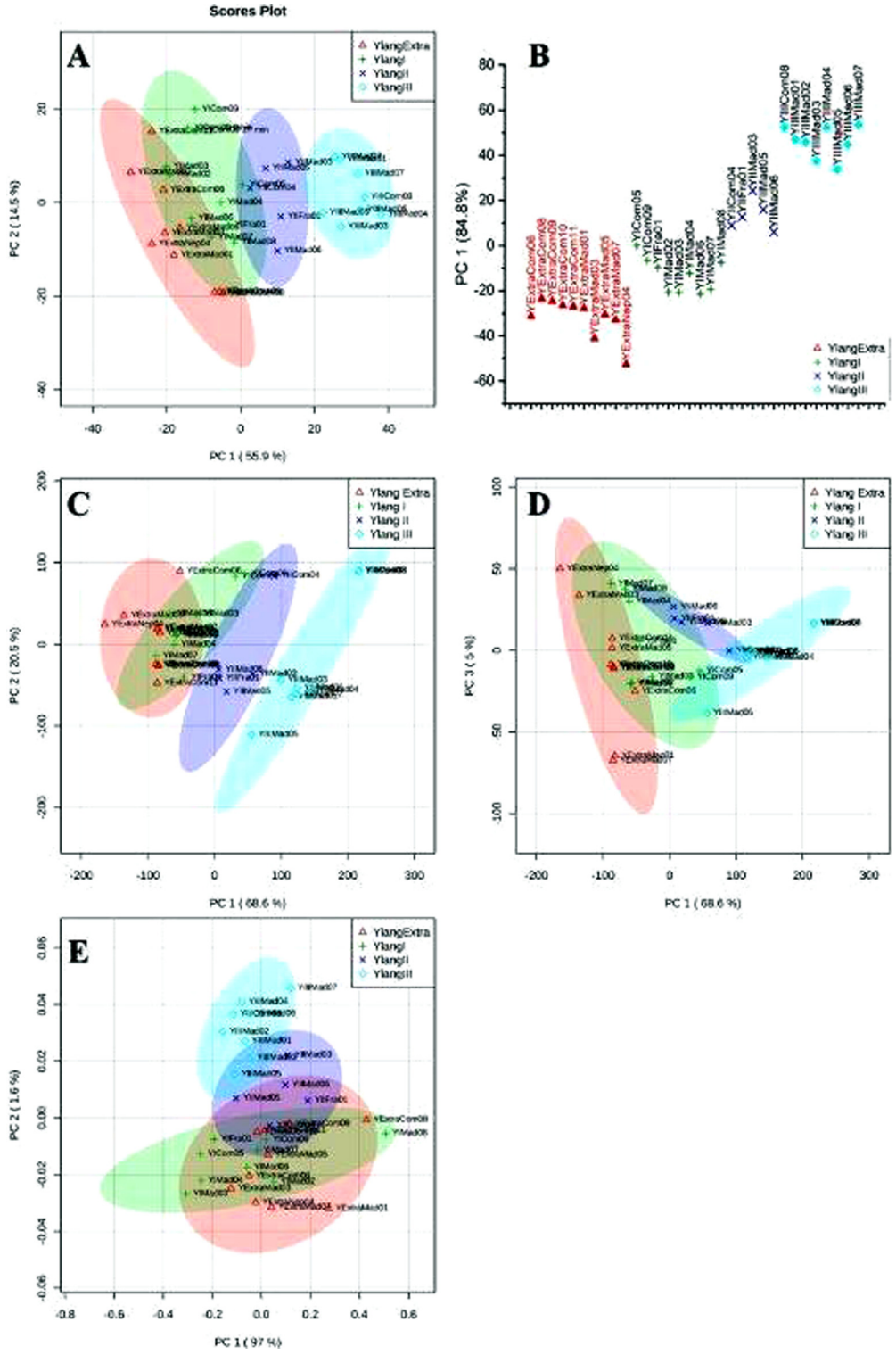

Fig. 2. Principal component analysis (PCA) of ylang-ylang essential oils according to grade (A) total chromatogram average mass spectrum (TCAMS) data, (B) segment average mass spectrum (SAMS) data, (C) principal components (PCs) 1 and 2 in the chemical composition data, (D) PCs 1 and 3 in the chemical composition data, and (E) attenuated total reflectance - Fourier transformation infrared spectroscopy (ATR-FTIR) data.

first, a dummy response matrix was created in which 1, 2, 3, and 4 corresponded to the grades Extra, I, II, and III respectively. Next, the optimum number of latent variables (LVs) required for the PLS model was investigated by observing the variance explained by each LV in the calibration and validation steps (Fig. S5). Cross-validation was performed in a leave-one-out manner. The optimal number of LVs for each of the models was determined by determining the last LV before calibration variance reached a plateau and there was an increase in the explained validation variance.

Predicted responses in the validation step were used to determine the grade of each sample, where values below 1.5 indicated the Extra grade, between 1.51–2.50 grade I, 2.51–3.50 grade II,

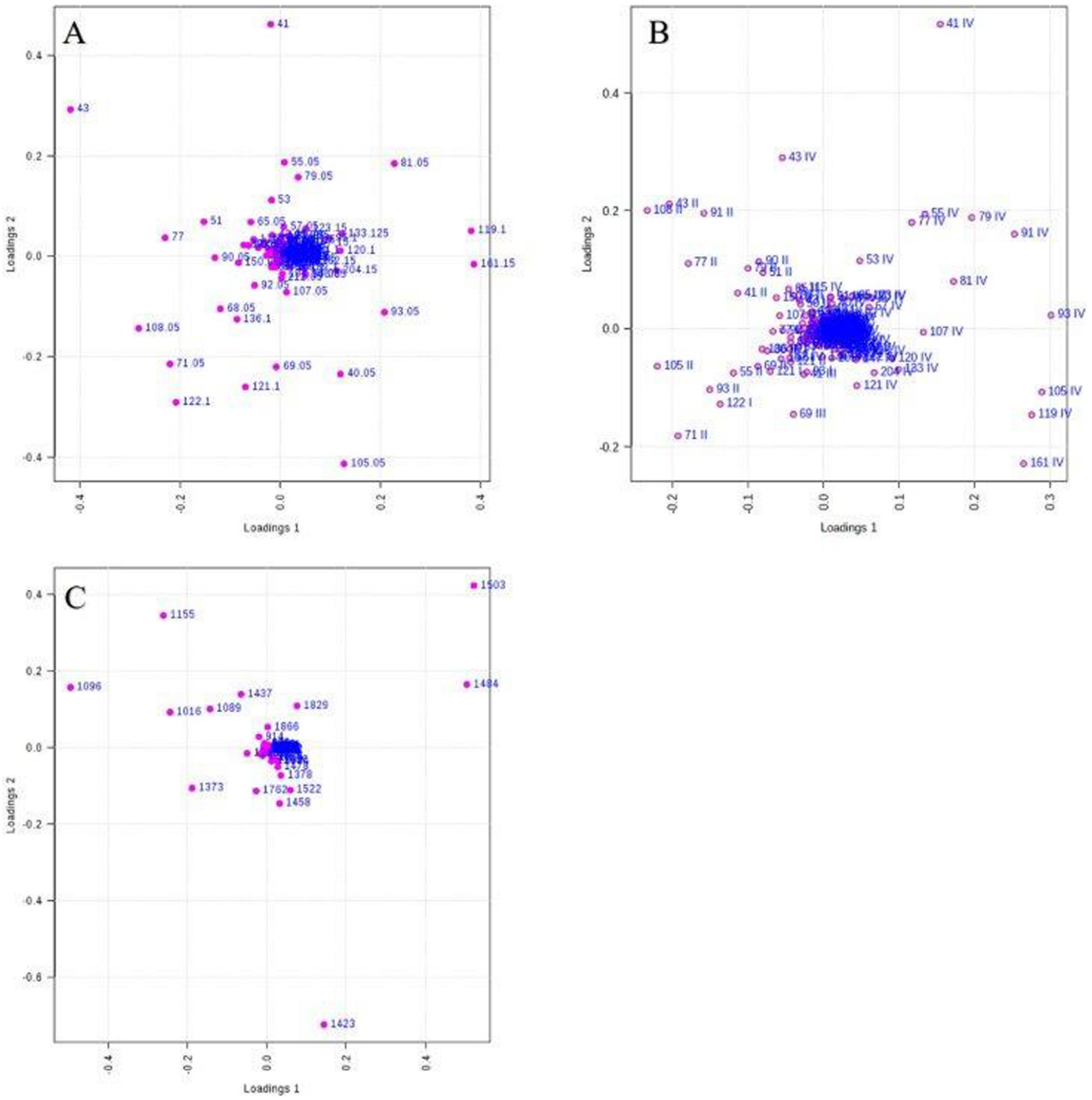

Fig. 3. Loadings plots for the principal component analysis (PCA) of the (A) total chromatogram average mass spectrum (TCAMS), (B) segment average mass spectrum (SAMS), and (C) Chemical profile data sets.

and 3.51–4.50 grade III. PLS prediction models, created on all 4 data sets, managed to differentiate between the YY EO grades, where only two samples, in both the chemical composition and ATR-FTIR data sets, were not classified in their appropriate grade, see Table 2 and Fig. S6. Bias for all models was 0, showing that there was no systematic error in any of the models.

The impact that each variable has on the prediction model was presented through weighted regression coefficients, where variables with positive coefficients were associated with higher grades, and variables with negative coefficients with lower grades. Three sesquiterpenes with retention indices 1423 (β-caryophyllene), 1484 (germacrene D) and 1503 ((*E,E*)-α-farnesene), were found to be most abundant in the higher grades, while lower, more fragrant grades had higher abundance in the 4 compounds: 1016 (*p*-cresyl methyl ether), 1089 (methyl benzoate), 1096 (linalool), 1155 (benzyl acetate), and 1437 (geranyl acetate) (see Fig. S7A).

In the PLS model built from the TCAMS data, the mass-fragments associated with the lower grades were *m/z* 43, 71, 108 and 122, which are characteristic for linalool, benzyl acetate and *p*-cresyl methyl ether, respectively. The higher grades were found to be richer in mass fragments 41, 81, 119 and 161 *m/z*, which arise in EI fragmentation of sesquiterpenes.

The same observations were made in evaluating the contribution of the variables to the PLS model from the SAMS data set.

Table 2
Measured and predicted values for the YY grades obtained by applying PLS regression models on total chromatogram average mass spectrum (TCAMS), segment average mass spectrum (SAMS), chemical composition, and attenuated total reflectance - Fourier transformation infrared spectroscopy (ATR-FTIR) data sets. Figures of merit for each model were the determined number of latent variables (LVs), root mean square error of calibration (RMSEC), root mean square error of cross-validation (RMSECV), slope, intercept, coefficient of determination (R^2) and bias.

	Reference	TCAMS Predicted	SAMS Predicted	Chemical composition Predicted	ATR-FTIR Predicted
YExtraMad01	1.00	1.04	1.34	0.85	1.26
YExtraMad03	1.00	1.08	0.87	1.05	1.18
YExtraCom06	1.00	1.31	1.21	1.03	1.17
YExtraMad07	1.00	0.96	1.16	0.76	1.33
YExtraMad05	1.00	1.41	1.24	1.26	1.40
YExtraNep04	1.00	0.95	0.46	0.93	1.66
YExtraCom08	1.00	1.13	1.49	1.32	1.18
YExtraCom09	1.00	1.10	1.45	1.33	1.28
YExtraCom10	1.00	1.02	1.38	1.31	1.31
YExtraCom11	1.00	1.46	1.38	1.63	1.33
YICom09	2.00	2.16	2.11	1.96	1.95
YICom05	2.00	2.45	2.34	2.17	2.33
YIMad02	2.00	1.52	1.61	1.57	1.71
YIFra01	2.00	2.10	2.00	2.14	1.76
YIMad03	2.00	1.54	1.60	1.79	1.84
YIMad06	2.00	1.57	1.57	1.55	1.54
YIMad07	2.00	1.81	1.65	1.91	1.61
YIMad08	2.00	2.04	2.08	1.75	1.59
YIMad04	2.00	2.15	1.91	2.07	1.96
YIIMad05	3.00	2.97	2.93	3.00	2.68
YIIFra01	3.00	2.95	2.81	2.85	2.44
YIIMad03	3.00	3.28	3.23	3.29	3.31
YIIMad06	3.00	2.62	2.56	2.88	2.46
YIICom04	3.00	2.67	2.67	2.85	2.53
YIIIMad07	4.00	4.21	4.26	4.16	4.71
YIIIMad02	4.00	3.98	4.00	4.08	3.95
YIIIMad01	4.00	4.04	4.04	3.97	3.79
YIIIMad05	4.00	3.54	3.56	3.40	3.54
YIIIMad06	4.00	4.05	3.96	4.04	3.88
YIIIMad03	4.00	3.70	3.69	3.61	3.55
YIIICom08	4.00	4.13	4.22	4.16	4.26
YIIIMad04	4.00	4.27	4.22	4.30	4.51
Number of LVs		2	1	3	3
RMSEC		0.27	0.31	0.28	0.39
RMSECV		0.29	0.31	0.32	0.44
Slope		0.95	0.93	0.95	0.90
Intercept		0.12	0.16	0.12	0.23
R^2		0.95	0.93	0.95	0.90
bias		0.00	0.00	0.00	0.00

The main compounds associated with grade discrimination were found in the second and fourth segments, see Fig. S7C. Analytes eluting in the second segment, such as *p*-cresyl methyl ether, methyl benzoate, linalool, and benzyl acetate, contributed mostly to the fragrance character of the lower YY grades. The higher abundance of heavier sesquiterpene compounds, eluting in the fourth segment, contribute to the less fragrant grades II and III.

3.3. Prediction of grade of ylang-ylang complete samples

Complete YY EOs are mixtures of all four distillation grades and therefore we used them as the validation sample set for our developed PLS grade model. Application of a PLS model based on the chemical profile did not result in a good grade prediction of Complete samples coming from Australia, Congo, and Indonesia (results not shown here). This result comes from the fact that there is additional variation in the chemical profiles based on the geographical origin of the samples, which was not contained in the PLS grade prediction model, since none of these origins were used in creating the model. Compared to the chemical composition data set, ATR-FTIR, TCAMS, and SAMS had a higher number of variables in the data array and managed to predict the grade of the samples from different geographical origins that were not used in creating the model. The PLS models based on the ATR-FTIR data set revealed that the Complete samples from Madagascar and Australia are the most similar to grade II, while samples from Indonesia, Congo and two samples from Madagascar are closer to the lower grades, see Fig. 4A.

However, as it can be observed from Fig. 4B and C, PLS models based on the TCAMS and SAMS data sets identified Indonesian, Australian and Madagascar samples as grade II. Samples YComplMad06 and YComplMad03 had lower predicted grades, closer to grade I. The YY Complete from Congo had a predicted grade lower than 1.5, which put it in the Extra grade. The lower predicted grade for the Congo samples can be connected with their higher abundances of highly volatile compounds, being more fragrant compared to Complete samples coming from Indonesia and Australia.

As it can be observed, there is significant difference between the uncertainty of prediction for the samples coming from Madagascar compared to the ones from Australia, Congo, and Indonesia. This comes from the fact that the samples in the training set were all from Madagascar and Comoros, while there were none from the other geographical origins. Nevertheless, the PLS models built were capable of determining the grades of the samples coming from other geographical locations, applying the same differences

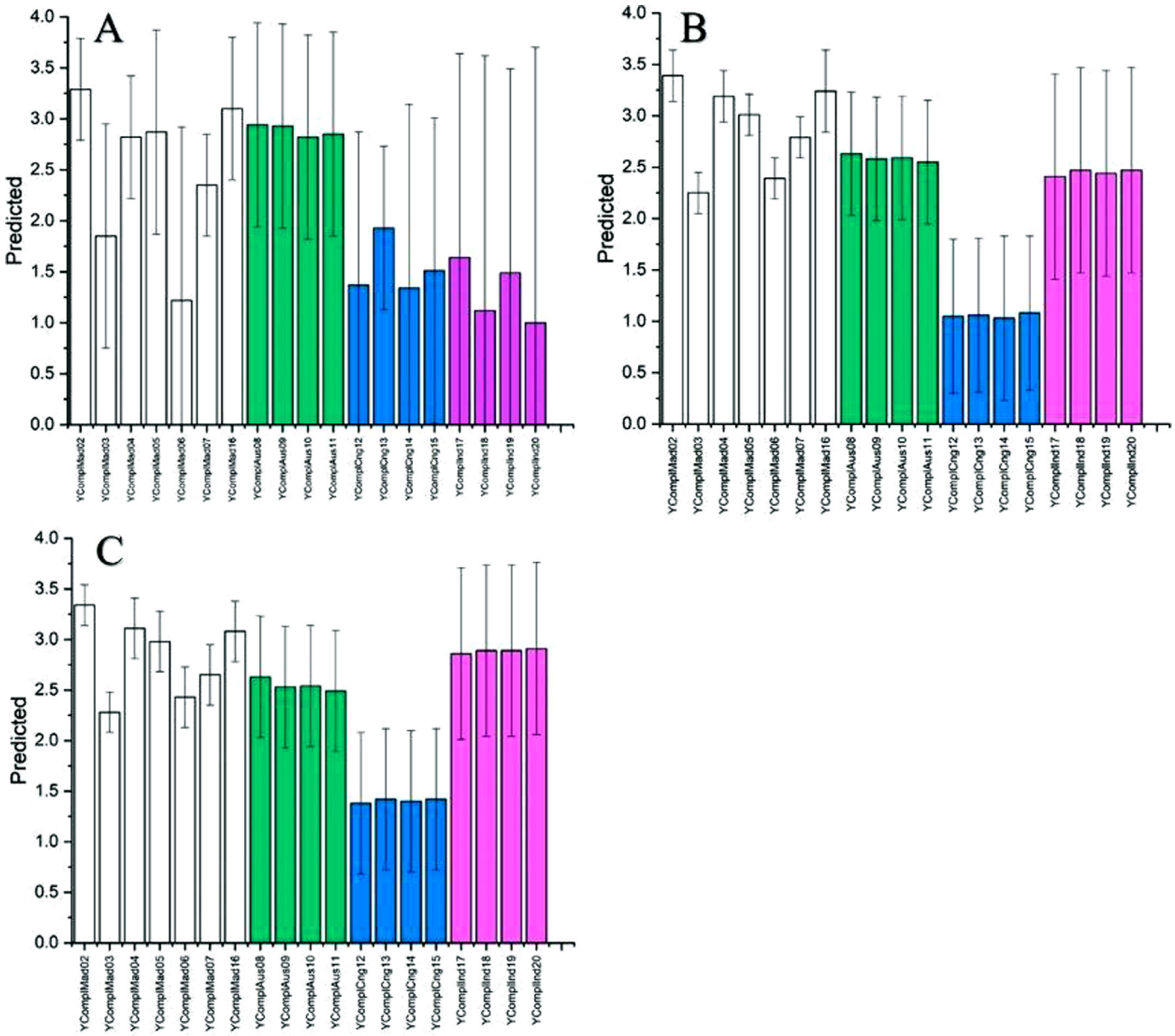

Fig. 4. Prediction of grade for the Complete samples coming from Madagascar, Australia, and Congo, and YY samples coming from Indonesia using the partial least squares (PLS) grade model on (A) attenuated total reflectance - Fourier transformation infrared spectroscopy (ATR-FTIR); (B) total chromatogram average mass spectrum (TCAMS), and (C) segment average mass spectrum (SAMS) data sets.

in the chemical profiles used in building the PLS grade models. Also, the uncertainty in the PLS models based on the TCAMS and SAMS data was significantly smaller than in the model created from the ATR-FTIR data set.

3.4. Discrimination of ylang-ylang EOs according to the geographical origin

The YY EO samples used in this study had several geographical origins (see Table 1). Most of the samples came from Madagascar and The Comoros, which are the biggest exporters of this EO. The Complete grade was obtained from Madagascar, Australia, and Congo, while YY EOs coming from Indonesia did not have a specified grade. Also included were two samples from France and one from Nepal. However, these samples, due to their small number, were not used in creating the prediction models.

In order to make comparisons between the geographical origins of the samples, it was necessary to avoid the influence of differences in distillation times. Since we did not have samples of all the grades from all geographical locations, a PLS-DA model was created from samples that belonged to the same grade. Accordingly, a single PLS-DA origin prediction model was created for the Complete samples coming from Madagascar, Australia, and Indonesia, while another was created for lower grade samples from Madagascar, Comoros, and Congo.

PLS-DA models based on TCAMS, SAMS and ATR-FTIR, differentiated between YY grade I EOs coming from Congo, Comoros, and Madagascar, as well as YY Complete samples coming from Australia, Indonesia and Madagascar. The model based on the chemical composition data set had sensitivity 0 for samples coming from Congo, as well as from Australia and Indonesia, which indicate that this model was not able to positively identify those

Table 3
Prediction of geographical origin (Congo, Comoros, Madagascar) and (Madagascar, Australia, Indonesia) using PLS-DA on total chromatogram average mass spectrum (TCAMS), segment average mass spectrum (SAMS), chemical composition and attenuated total reflectance - Fourier transformation infrared spectroscopy (ATR-FTIR) data sets. Specificity and sensitivity were determined for both calibration (Cal) and cross-validation (CV) sample sets, as well as number of latent variables (LV) used in creating the PLS-DA model.

		TCAMS		SAMS		Chemical composition		ATR-FTIR	
		Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
Comoros	**Cal**	1.0	1.0	1.0	1.0	0.60	1.00	1.00	1.00
	CV	1.0	1.0	1.0	1.0	0.60	0.86	1.00	1.00
Congo	**Cal**	1.0	1.0	1.0	1.0	1.00	0.00	1.00	1.00
	CV	1.0	1.0	1.0	1.0	1.00	0.00	1.00	1.00
Madagascar	**Cal**	1.0	1.0	1.0	1.0	1.00	1.00	1.00	1.00
	CV	1.0	1.0	1.0	1.0	0.91	1.00	1.00	1.00
Number of LVs		6		5		2		5	
Explained variance in CV (%)		88.52		89.96		38.15		85.47	
Australia	**Cal**	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00
	CV	1.00	1.00	1.00	1.00	0.91	0.00	1.00	1.00
Indonesia	**Cal**	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00
	CV	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00
Madagascar	**Cal**	1.00	1.00	1.00	1.00	1.00	0.86	1.00	1.00
	CV	1.00	1.00	1.00	1.00	1.00	0.86	1.00	1.00
Number of LVs		2		2		1		3	
Explained variance in CV (%)		98.81		98.1		29.47		89.56	

samples. Observing the parameters of prediction (see Table 3), SAMS required the least number of LVs, as well as the explained variance was the highest, compared to other data sets which differentiated between all origins.

3.5. Prediction of chemical composition

Since the abundances of the *m/z* fragments are directly dependent on the qualitative and quantitative characteristics of the analysed samples, it was possible to find mathematical models to calculate the relative abundances of the analytes based on the measured *m/z* abundances. In this way, a PLS prediction model for the quantification of 18 compounds based on GC–FID relative abundances, was created. ATR-FTIR coupled with PLS has already been established in the literature as the method that can be used for prediction of chemical composition of EOs [21,22], and it will be used as a benchmark method. Evaluating the prediction model parameters, TCAMS and ATR-FTIR showed very similar performances, where in the case of prenyl acetate, (*E*)-cinnamyl aldehyde, methyl benzoate, and (*E,E*)-farnesyl acetate, ATR-FTIR showed better performance, while in case of β-caryophyllene, γ-muurolene, and δ-cadinene, TCAMS had a lower REP (Fig. S8).

SAMS showed equal or better prediction performance (Table 4), compared to the other data sets, in prediction of chemical composition, where 9 out of 18 compounds were predicted with a REP lower than 10%, six with REP values between 10-15%, and only three having REP values above 15% (see Fig. S8).

3.6. Application of PLS prediction models in quality assessment of YY EOs

Quality assurance of different distillation grades of YY EOs is done by comparing the abundance of the main compounds with the available ISO 3063:2004 standard "Oil of ylang-ylang [*Cananga odorata* (Lam.) Hook. f Thomson forma *genuine*]". Among the 18 predicted compounds, there were 15 that are used to assess the compliance of YY samples with the ISO 3063:2004 standard. Here, we explore the possibility of using the values predicted by PLS prediction models from TCAMS, SAMS, and ATR-FTIR data in the quality assessment of YY EOs. Misclassification (MC) and REP were used as the parameter for assessing the goodness of application of the PLS models for quality assessment. MC can be defined as a percentage of false positives and negatives compared to total number of comparisons.

In the case of the PLS model created on the SAMS data set, the calculated MC was 2.5%, while in the case of the TCAMS ATR-FTIR data sets this value was 6%. Most of the false positives and negatives were with the values close to the range limits where even small differences led to different conclusions. Also, compounds with small abundances, such as prenyl acetate, had higher REP values, which led to higher rates of misclassification.

4. Conclusion

This study outlines the advantages of utilising a *m/z* fragment-based chemometric approach for characterisation of YY EOs, as compared to traditional procedures, chemical composition, and ATR-FTIR. Performances of two GC–qMS-based data sets (TCAMS and SAMS) in EO discrimination between grades, origins, and in quality assessment, were evaluated using PCA and PLS models. It was clear that TCAMS-based models are much faster and more readily created. The only requirement to apply SAMS with such statistical techniques is to match the segments between the samples, where the same compounds must be picked in the same segment through the whole data set.

PCA performed on the SAMS data set required only PC 1, while in case of TCAMS it required two PCs to reveal the trend between different YY grades.

In this work, it was shown that PLS regression models are a fast and robust approach for the evaluation of YY EO grades (fragrance), which is of immense importance in the perfume and cosmetic industries. Also, using PLS-DA models, it was possible to differentiate between YY picking locations/geographic origins.

Quantitative characterisation of 18 compounds was successfully performed creating PLS regression models on TCAMS and SAMS data sets, eliminating the quantification by GC–FID. The PLS model created on the SAMS data showed superior performance with a REP lower than 10% for most compounds. The values obtained were used to evaluate the quality of the YY samples in comparison with the appropriate ISO standard. MC values of 6% and 2.5% for the TCAMS and SAMS data sets shows possibility for their application in EO quality assessment.

In this work, the potential for application of two readily created data sets from GC–MS raw data, TCAMS and SAMS, in YY grade and origin was demonstrated. Also, quality control of YY EO based on TCAMS and SAMS have clear advantages over traditional approach.

Table 4
Figures of merit for PLS regression models of 18 compounds in the YY EO using total chromatogram average mass spectrum (TCAMS), segment average mass spectrum (SAMS) and attenuated total reflectance - Fourier transformation infrared spectroscopy (ATR-FTIR) data. Figures of merit consist of the number of latent variables (LVs), root mean square error of calibration (RMSEC), root mean square error of cross-validation (RMSECV), relative error of prediction (REP) and coefficient of determination (R^2).

		TCAMS						SAMS						ATR-FTIR					
	Concentration range (%)	LVs	RMSEC	RMSECV	REP(%)	R^2	Bias	LVs	RMSEC	RMSECV	REP(%)	R^2	Bias	LVs	RMSEC	RMSECV	REP(%)	R^2	Bias
Prenyl acetate	0.00–0.97	4	0.16	0.18	69.7	0.70	0.0	8	0.06	0.10	42.4	0.96	0.0	7	0.06	0.09	42.2	0.94	0.0
***p*-Cresyl methyl ether**	0.11–7.58	10	0.50	0.65	18.1	0.97	0.0	8	0.34	0.52	14.2	0.98	0.0	10	0.33	0.65	17.6	0.98	0.0
Methyl benzoate	0.00–5.70	9	0.34	0.43	18.4	0.98	0.0	6	0.14	0.17	7.5	0.99	0.0	10	0.21	0.33	14.4	0.99	0.0
Linalool	0.27–20.53	9	0.68	0.98	13.8	0.98	0.0	10	0.34	0.64	8.7	0.99	0.0	10	0.69	1.17	15.7	0.98	0.0
Benzyl acetate	0.00–12.05	8	0.69	0.91	20.0	0.97	0.0	10	0.18	0.33	7.0	0.99	0.0	9	0.46	0.87	18.4	0.98	0.0
Geraniol	0.00–2.67	6	0.19	0.23	28.8	0.90	0.0	10	0.09	0.18	20.8	0.98	0.0	10	0.17	0.24	27.6	0.95	0.0
Geranyl acetate	0.81–9.82	8	0.64	0.85	17.1	0.94	0.0	8	0.25	0.38	7.7	0.99	0.0	10	0.37	0.61	12.1	0.98	0.0
β-Caryophyllene	4.40–60.86	9	1.05	1.51	8.6	0.99	0.0	10	0.59	1.09	6.2	0.99	0.0	10	1.37	2.46	13.9	0.99	0.0
(*E*)-Cinnamyl aldehyde	0.00–4.63	7	0.38	0.49	34.0	0.92	0.0	10	0.11	0.18	12.4	0.99	0.0	10	0.19	0.34	23.7	0.98	0.0
α-Humulene	1.43–6.30	9	0.40	0.52	15.7	0.91	0.0	8	0.21	0.36	10.7	0.98	0.0	10	0.22	0.37	11.1	0.98	0.0
γ-Muurolene	0.45–2.34	5	0.18	0.20	15.5	0.89	0.0	7	0.11	0.16	12.4	0.96	0.0	10	0.14	0.23	17.7	0.94	0.0
Germacrene D	1.10–24.96	8	1.31	1.73	14.1	0.96	0.0	10	0.54	1.00	8.4	0.99	0.0	9	0.98	1.41	11.8	0.98	0.0
(*E,E*)-α-Farnesene	0.73–25.13	9	0.75	1.23	15.4	0.98	0.0	10	0.41	0.74	9.5	0.99	0.0	10	0.82	1.22	15.2	0.98	0.0
δ-Cadinene	0.24–4.87	6	0.34	0.38	14.6	0.91	0.0	8	0.17	0.28	10.5	0.98	0.0	10	0.25	0.45	17.2	0.96	0.0
(*E,E*)-Farnesol	0.27–2.54	8	0.19	0.25	17.9	0.92	0.0	10	0.07	0.13	9.4	0.99	0.0	7	0.18	0.24	16.9	0.92	0.0
Benzyl benzoate	4.88–14.02	9	0.40	0.60	8.2	0.97	0.0	10	0.26	0.43	5.8	0.99	0.0	10	0.40	0.60	8.1	0.97	0.0
(*E,E*)-Farnesyl acetate	0.09–5.48	6	0.36	0.45	25.8	0.92	0.0	8	0.21	0.34	20.4	0.97	0.0	10	0.22	0.33	19.9	0.97	0.0
Benzyl salicylate	0.79–5.87	10	0.25	0.38	13.4	0.96	0.0	10	0.16	0.30	10.5	0.99	0.0	10	0.24	0.38	13.3	0.97	0.0

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Leo Lebanov: Investigation, Conceptualization, Methodology, Writing - original draft. **Shiladitya Chatterjee:** Methodology, Writing - review & editing. **Laura Tedone:** Methodology, Investigation. **Sean C. Chapman:** Methodology. **Matthew R. Linford:** Conceptualization, Methodology, Writing - review & editing, Software, Validation. **Brett Paull:** Conceptualization, Methodology, Writing - review & editing.

Acknowledgments

The authors would like to thank the University of Tasmania, Australia, and Plant Therapy Inc, Twin Falls, ID, USA, for their generous provision of a co-funded PhD scholarship and project support for Leo Lebanov. Grant number: IH170100020. The authors would also like to thank Ahimsa Oils, Australia and Albert Vieille SAS, France for the essential oil samples.

Ethical standards

Authors declare that they have no conflict of interest.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.chroma.2020.460853.

References

[1] L. Lebanov, L. Tedone, M. Kaykhaii, M.R. Linford, B. Paull, Multidimensional gas chromatography in essential oil analysis. Part 2: application to characterisation and identification, Chromatographia 82 (2019) 399–414, doi:10.1007/s10337-018-3651-9.
[2] P. Rubiolo, B. Sgorbini, E. Liberto, C. Cordero, C. Bicchi, Essential oils and volatiles: sample preparation and analysis. A review, Flavour Fragr. J. 25 (2010) 282–290, doi:10.1002/ffj.1984.
[3] P.H. Stefanuto, K.A. Perrault, L.M. Dubois, B. L'Homme, C. Allen, C. Loughnane, N. Ochiai, J.F. Focant, Advanced method optimization for volatile aroma profiling of beer using two-dimensional gas chromatography time-of-flight mass spectrometry, J. Chromatogr. A 1507 (2017) 45–52, doi:10.1016/j.chroma.2017.05.064.
[4] L. Lebanov, L. Tedone, M. Kaykhaii, M.R. Linford, B. Paull, Multidimensional gas chromatography in essential oil analysis. Part 1: technical developments, Chromatographia 82 (2019) 377–398, doi:10.1007/s10337-018-3649-3.
[5] M. Heel, R.V. Portugal, M. Schatz, Multivariate statistical analysis of large datasets: single particle electron microscopy, Open J. Stat. 6 (2016) 701–739, doi:10.4236/ojs.2016.64059.
[6] Z. Xiang, X.Q. Wang, X.J. Cai, S. Zeng, Metabolomics study on quality control and discrimination of three curcuma species based on gas chromatograph-mass spectrometry, Phytochem. Anal. 22 (2011) 411–418, doi:10.1002/pca.1296.
[7] G. Flamini, M. Tebano, P.L. Cioni, Y. Bagci, H. Dural, K. Ertugrul, T. Uysal, A. Savran, A multivariate statistical approach to Centaurea classification using essential oil composition data of some species from Turkey, Plant Syst. Evol. 261 (2006) 217–228, doi:10.1007/s00606-006-0448-3.
[8] K.H. Liland, Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis, TrAC, Trends Anal. Chem. 30 (2011) 827–841, doi:10.1016/j.trac.2011.02.007.
[9] T.B. Yang, P. Yan, M. He, L. Hong, R. Pei, Z.M. Zhang, L.Z. Yi, X.Y. Yuan, Application of subwindow factor analysis and mass spectral information for accurate alignment of non-targeted metabolic profiling, J. Chromatogr. A 1563 (2018) 162–170, doi:10.1016/j.chroma.2018.05.071.

[10] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 (2006) 779–787, doi:10.1021/ac051437y.
[11] T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinform. 11 (2010) 395, doi:10.1186/1471-2105-11-395.
[12] E. Lange, R. Tautenhahn, S. Neumann, C. Gropl, Critical assessment of alignment procedures for LC–MS proteomics and metabolomics measurements, BMC Bioinform. 9 (2008) 375, doi:10.1186/1471-2105-9-375.
[13] N.S. Radulović, P.D. Blagojević, Average mass scan of the total ion chromatograms: a new gas chromatography-mass spectrometry derived variable for fast and reliable multivariate statistical treatment of essential oil compositional data, J. Chromatogr. A 1301 (2013) 190–199, doi:10.1016/j.chroma.2013.05.039.
[14] L.A. Adutwum, R.J. Abel, J. Harynuk, Total ion spectra versus segmented total ion spectra as preprocessing tools for gas chromatography - mass spectrometry data, J. Forensic Sci. (2017), doi:10.1111/1556-4029.13657.
[15] J. Vestner, G. de Revel, S. Krieger-Weber, D. Rauhut, M. du Toit, A. de Villiers, Toward automated chromatographic fingerprinting: a non-alignment approach to gas chromatography mass spectrometry data, Anal. Chim. Acta 911 (2016) 42–58, doi:10.1016/j.aca.2016.01.020.
[16] R. Bro, Review on multiway analysis in chemistry - 2000-2005, Crit. Rev. Anal. Chem. 36 (2006) 279–293, doi:10.1080/10408340600969965.
[17] L.G. Johnsen, P.B. Skou, B. Khakimov, R. Bro, Gas chromatography - mass spectrometry data processing made easy, J. Chromatogr. A 1503 (2017) 57–64, doi:10.1016/j.chroma.2017.04.052.
[18] P. Ma, Z. Zhang, X. Zhou, Y. Yun, Y. Liang, H. Lu, Feature extraction from resolution perspective for gas chromatography-mass spectrometry datasets, RSC Adv. 6 (2016) 113997–114004, doi:10.1039/c6ra17864b.
[19] M.E. Sigman, M.R. Williams, J.A. Castelbuono, J.G. Colca, C.D. Clark, Ignitable liquid classification and identification using the summed-ion mass spectrum, Instrum. Sci. Technol. 36 (2008) 375–393, doi:10.1080/10739140802151440.
[20] L. Lebanov, L. Tedone, A. Ghiasvand, B. Paull, Random Forests machine learning applied to gas chromatography – mass spectrometry derived average mass spectrum data sets for classification and characterisation of essential oils, Talanta 208 (2020), doi:10.1016/j.talanta.2019.120471.
[21] I. Bombarda, N. Dupuy, J.P. Da, E.M. Gaydou, Comparative chemometric analyses of geographic origins and compositions of lavandin var. Grosso essential oils by mid infrared spectroscopy and gas chromatography, Anal. Chim. Acta 613 (2008) 31–39, doi:10.1016/j.aca.2008.02.038.
[22] S. Lafhal, P. Vanloot, I. Bombarda, J. Kister, N. Dupuy, Identification of metabolomic markers of lavender and lavandin essential oils using mid-infrared spectroscopy, Vib. Spectrosc. 85 (2016) 79–90, doi:10.1016/j.vibspec.2016.04.004.
[23] C. Benini, M. Ringuet, J.-P. Wathelet, G. Lognay, P. Jardin, M.-L. Fauconnier, Variations in the essential oils from ylang-ylang (*Cananga odorata* [Lam.] Hook f. & Thomson forma *genuina*) in the Western Indian Ocean islands, Flavour Fragr. J. 27 (2012) 356–366, doi:10.1002/ffj.3106.
[24] M. Brokl, M.L. Fauconnier, C. Benini, G. Lognay, P. du Jardin, J.F. Focant, Improvement of ylang-ylang essential oil characterization by GCxGC-TOFMS, Molecules 18 (2013) 1783–1797, doi:10.3390/molecules18021783.
[25] H. van Den Dool, P. Dec, Kratz, A generalization of the retention index system including linear temperature programmed gas—liquid partition chromatography, J. Chromatogr. A 11 (1963) 463–471, doi:10.1016/S0021-9673(01)80947-X.
[26] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D.S. Wishart, J. Xia, MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis, Nucl. Acids Res. 46 (2018) 486–494, doi:10.1093/nar/gky310.
[27] S. Kucheryavskiy, S. R Package 'mdatools': Multivariate Data Analysis for Chemometrics, R package version 0.8. 2, CRAN, 2017.
[28] S. Chatterjee, B. Singh, A. Diwan, Z.R. Lee, M.H. Engelhard, J. Terry, H.D. Tolley, N.B. Gallagher, M.R. Linford, A perspective on two chemometrics tools: PCA and MCR, and introduction of a new one: Pattern recognition entropy (PRE), as applied to XPS and ToF-SIMS depth profiles of organic and inorganic materials, Appl. Surf. Sci. 433 (2018) 994–1017, doi:10.1016/j.apsusc.2017.09.210.
[29] J.E. Jackson, A User's Guide To Principal Components, John Wiley & Sons, Inc., New York, 1991.
[30] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831, doi:10.1039/c3ay41907j.
[31] B.N. Nsuala, G.P. Kamatou, M. Sandasi, G. Enslin, A. Viljoen, Variation in essential oil composition of Leonotis leonurus, an important medicinal plant in South Africa, Biochem. Syst. Ecol. 70 (2017) 155–161, doi:10.1016/j.bse.2016.11.009.
[32] T. Rajalahti, O.M. Kvalheim, Multivariate data analysis in pharmaceutics: a tutorial review, Int. J. Pharm. 417 (2011) 280–290, doi:10.1016/j.ijpharm.2011.02.019.
[33] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130, doi:10.1016/S0169-7439(01)00155-1.
[34] D.I. Broadhurst, D.B. Kell, Statistical strategies for avoiding false discoveries in metabolomics and related experiments, Metabolomics 2 (2006) 171–196, doi:10.1007/s11306-006-0037-z.
[35] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Assessment of PLSDA cross validation, Metabolomics 4 (2008) 81–89, doi:10.1007/s11306-007-0099-6.
[36] E. Szymanska, E. Saccenti, A.K. Smilde, J.A. Westerhuis, Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics 8 (2012) 3–16, doi:10.1007/s11306-011-0330-3.
[37] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods (2013) 5, doi:10.1039/c3ay40582f.
[38] ISO 3063, Oil of Ylang-Ylang (Cananga odorata (Lam.) Hook. F. et Thomson forma Genuina) ©ISO 2004; Geneva, Switzerland, 2004 www.iso.org.
[39] Y.J. Liu, T. Tran, G. Postma, L.M.C. Buydens, J. Jansen, Estimating the number of components and detecting outliers using Angle Distribution of Loading Subspaces (ADLS) in PCA analysis, Anal. Chim. Acta 1020 (2018) 17–29, doi:10.1016/j.aca.2018.03.044.