Fuel 327 (2022) 124715

Contents lists available at ScienceDirect

Fuel

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

Full Length Article

An original deconvolution approach for oil production allocation based on geochemical fingerprinting

Leonardo Sandoval [a], Monica Riva [a], Placido Franco [b], Ivo Colombo [b], Roberto Galimberti [b], Alberto Guadagnini [a,*]

[a] Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Piazza Leonardo Da Vinci 32, Milan 20133, Italy
[b] Geolog Technologies, Via Monte Nero 30, San Giuliano Milanese 20098, Italy

ARTICLE INFO

Keywords:
Reservoir geochemistry
Production allocation
Geochemical fingerprinting
Maximum Likelihood
GC–MS

ABSTRACT

We tackle oil commingling scenarios and develop an original deconvolution approach for geochemical production allocation. This yields robust assessment of the proportions of oils forming a mixture originating from commingling oils associated with diverse reservoirs or, wells. Our study starts from considering that production allocation performed by means of geochemical fingerprinting is relevant in the context of modern and sustainable use of georesources, with the added benefit of favoring shared facilities and production equipment. A geochemical production allocation workflow is typically structured according to two steps: (i) determination of the chromatograms associated with the mixture (and eventually with each of the End Members, EMs, constituting the fluids in the mixture), and (ii) the use of a deconvolution algorithm to estimate the mass fraction of each EM. Concerning the latter step, we introduce an original approach and the ensuing deconvolution algorithm (hereafter termed PGM) that does not require additional laboratory efforts in comparison with traditional approaches. We also present extensions of widely used deconvolution algorithms, which we frame in a (stochastic) Monte Carlo context to improve their robustness and reliability. The new PGM approach is assessed jointly with a suite of typically used approaches and algorithms against new laboratory-based commingling scenarios. The latter are based on the design and introduction of a novel and low-cost experimental method. The results of the study (i) constitute a unique and rigorous comparison of the traditionally employed production allocation deconvolution algorithms, (ii) document the critical importance of the number of features of the chromatograms used during a quantitative deconvolution, and (iii) suggest that our new PGM approach is very robust and accurate compared to existing approaches.

1. Introduction

The assessment of the spatial and temporal chemical evolution of fluids in oil and gas systems is a key element of modern reservoir engineering and use of underground energy resources. It provides strong support to the planning and implementation of effective and sustainable strategies for reservoir development and production [1]. In this context, interpretations arising from geochemical analyses can have significant impacts to cost and production scenarios. Advancements in geochemical tools and related analytical methodologies enable further constraining of uncertainties associated with reservoir characterization [2]. Within this broad framework, geochemical fingerprinting is nowadays considered a robust technique in reservoir geochemistry applications [3–7]. Geochemical analyses of reservoir fluids are routinely adopted to support our conceptual understanding of fluid connectivity. The latter represents a major aspect in reservoir modeling and management and can markedly impact oil and gas production [8,9]. Additionally, geochemical data can be used for dynamic reservoir performance assessment and evaluation of compositionally graded fluid column depletion [10]. The adoption of such techniques yields robust results, which are overall consistent with findings obtained with other classical methods (such as, e.g., pressure tests or wireline logs for compartmentalization assessment, as well as production logging tools and flowmeters for production monitoring) and are characterized by marked advantages in terms of cost savings and practical convenience [11–13].

Oil commingling is a common practice in the petroleum/georesources industry. It is associated with the benefit of cost reduction by sharing facilities and production equipment. Crude oils originated from

* Corresponding author.
E-mail address: alberto.guadagnini@polimi.it (A. Guadagnini).

https://doi.org/10.1016/j.fuel.2022.124715
Received 4 March 2022; Received in revised form 11 May 2022; Accepted 28 May 2022
Available online 4 July 2022

various reservoirs, wells, and/or fields are mixed and jointly produced through commingling operations. It may be necessary to deconvolute such mixtures, with the goal of assessing the individually contributing zones in the subsurface. The process of assigning the individual contribution of an oil type to the overall production is known as *production allocation* and has been subject to renewed interest in recent years [3,14–17].

Conventional production allocation techniques [18–20] make use of the molecular differences between individual End Members, EMs, i.e., fluids belonging to a distinct region in the system. In production allocation scenarios, commingled oils and EMs are typically analyzed through geochemical techniques (e.g., gas chromatography, GC). The ensuing data (chromatograms, also termed GC fingerprints) are then processed upon relying on deconvolution algorithms to evaluate the contribution of each EM to the commingled produced oil. Several production allocation methodologies and procedures have been illustrated in the literature. Key differences among them are chiefly related to the experimental setup for GC fingerprinting and the deconvolution algorithm employed to analyze the ensuing data.

Three main GC experimental approaches can be identified: (i) High-resolution gas chromatography (HRGC), targeting the aliphatic and aromatic peaks lying between dominant n-alkanes; (ii) Multidimensional gas chromatography (MDGC) based on the quantitative target analysis of C_8-C_9 alkylbenzenes; and (iii) Saturate and aromatic fraction gas chromatography-mass spectrometry (GC–MS) analysis.

HRGC [20,21] is based on the GC analysis of whole-oil samples. The chemical species considered are typically in the range C_8-C_{20} and the acquired chromatograms provide peak heights of a variety (sometimes hundreds) of compounds (which mostly remain unidentified). The main element guiding peak selection for the quantitative deconvolution is the possibility of finding a set of components that can enable one to discriminate between EMs and can therefore be used to deconvolute commingled oil samples. Otherwise, peak selection can be somehow arbitrary and largely a subjective (and operator-dependent) element. Moreover, HRGC is often plagued by poor chromatographic resolution. Accuracy can then deteriorate with time, as it might be significantly affected by changes in detector response and baseline drift.

MDGC [22] focuses on a limited number of compounds (11 alkylbenzenes). These can be considered as a constrained, while representative, dataset capable of explaining much of the variability between oil groups. The MDGC technique selectively detects a limited number of compounds, all of them chromatographically well separated, a feature which positively affects the accuracy, repeatability, and reproducibility of the analysis. Therefore, peak heights can be readily determined, since either external or internal standard calibrations can be conveniently carried out on a constrained number of components. This element is markedly relevant when analyses are performed across wide temporal windows (e.g., during production monitoring activities). In such cases, newly assimilated samples can be analyzed over time without the need to re-analyze previously acquired data.

Finally, GC–MS analyses [23,24] have the notable advantage of unambiguously identifying an extensive set of geochemical features in the chromatogram. The resulting database can then be used to identify basin-specific indicators which are directly linked to up-to-date production/performance data. This approach might require that multiple analyses (one for each type of component) be carried out on the same sample. As such, this makes (in principle) the methodology significantly more expensive and time-consuming than MDGC. Furthermore, some of these analyses require pre-analytical steps (for example fractionation through open column/medium pressure GC). These can introduce further biases, thus potentially affecting the accuracy of the deconvolution results.

When considering deconvolution algorithms for data processing, a variety of approaches have been described in the literature [22–29]. These approaches can be framed within the general context of system identification, also leveraging on statistical signal processes analysis [30]. In this framework, a system is excited by one or multiple known signals (e.g., EM chromatograms). The objective is then to estimate the target response (i.e., EMs mass fraction in the mixture) from available observations of the system output (i.e., mixture chromatogram). It is noted that system identification theory is used within a wide range of areas. These encompass, e.g., structural system analysis, medical image processing, wireless communication systems, and Earth sciences (e.g., [31] and references therein). For instance, it has been applied in oil exploration engineering to estimate properties of geomaterials based on wave signal data [32].

Deconvolution algorithms can be grouped into two main categories (i) methods based on peak heights (or actual concentrations) and (ii) methods based on peak ratios, evaluated from peaks associated with molecules eluting at close times. The use of peak ratios instead of peak heights is consistent with the possible change of baseline of the GC detection due to the use of multiple GC devices or improper equipment calibration. Note that a baseline change can significantly affect the monitored peak height whereas the peak ratios remain unaltered [33].

The first algorithm for geochemical production allocation has been proposed by Kaufman et al. [20], who exploited peak ratios in the C_{15}-C_{20} molecular range. This approach considers the fractional composition of an EM in a mixture to be related to the difference between the peak ratios of the EM and of the mixture. As highlighted by McCaffrey et al. [29], the approach suffers from two main drawbacks: (i) ratios of mixture chromatogram are not linear combinations of the ratios of EMs, so that the use of artificial mixtures with known EM contributions is required; (ii) the method is typically restricted to the allocation of mixtures composed by (at most) three EMs, as the mixing curves are not associated with a simple graphical representation. McCaffrey et al. [29] proposed an approach that makes use of peak heights (rather than peak ratios) where the relative amount of each compound/molecule (that is proportional to peak heights) in commingled samples is the result of a (weighted) linear combination of the concentration of molecules in each of the EMs. Relying on this approach circumvents the need for artificial mixtures, and deconvolution is not limited to the aforementioned three EMs. The approach introduced by Nouvelle et al. [28] is based on peak ratios and is aimed at (i) overcoming errors associated with baseline change of GC and (ii) enabling production allocation for mixtures with virtually no limitation on the number of EMs. Finally, recent efforts have been directed towards the development of approaches conducive to production allocation without strictly requiring information on EMs. A promising method is based on the Alternating Least Squares (ALS) algorithm [25,34].

In this broad context, the distinctive aim of our study is to introduce an original deconvolution algorithm that (i) makes use of peak ratios, thus providing high flexibility against possible sampling errors caused by baselines changes and/or improper equipment calibration and (ii) does not require relying on synthetic mixtures, thus reducing efforts and resources, in terms of laboratory time and investments. We do so upon framing our methodology within a technical and theoretical assessment of the deconvolution approaches discussed above and including extensions of (a) the method proposed by Nouvelle et al. [28] and (b) the ALS algorithm, which we view in a stochastic context. All of these approaches are then considered against a suite of new laboratory-based commingling scenarios. Concerning these, we also introduce a novel and low-cost GC experimental approach. The latter is grounded on a direct quantitative determination of C_8-C_{12} alkylbenzene components in oil through GC–MS fingerprinting and has been developed to circumvent some limitations of the typically employed methodologies.

2. Methods

Here, we briefly introduce in Sect. 2.1 two mixing models which are traditionally used for production allocation and rely on peak heights and peak height ratios. We then present an appraisal of key elements of two widely used deconvolution algorithms (Sect. 2.2), including their area of application, advantages, and limitations. Sect. 2.3 is devoted to the introduction of an original deconvolution algorithm that overcomes some of the limitations detected in the traditional approaches. We conclude the analysis by discussing (in Sect. 2.4) the ALS algorithm. The latter is used for the deconvolution of commingled fluids in cases where the absence of information on EMs hampers the applicability of the previously considered approaches. Here, we propose to extend ALS by viewing it in the context of a stochastic (Monte Carlo) framework.

2.1. *Mixing models*

Approaches to deconvolution in the context of production allocation are grounded on mixing models of either peak heights (or peaks) of a chromatogram or peak height ratios (hereafter termed ratios) evaluated from peaks associated with molecules eluting at close times.

Mixing models associated with peak heights rest on the assumption that peaks in the GC of a mixture are linear combinations of peaks of the GCs associated with each EM according to

$$\underline{\underline{\mathbf{A}}}\mathbf{x} = \mathbf{b} + \boldsymbol{\varepsilon} \qquad \text{with} \qquad \sum_{k=1}^{K} x_k = 1, \tag{1}$$

where, $\mathbf{x} = (x_1, \cdots, x_K)^{\mathrm{T}}$ is a vector containing the (unknown) mass fractions (x_k; $k = 1,\ldots,K$) of the K EMs in a mixture; $\mathbf{b} = \left(b_1, \cdots, b_{N_p}\right)^{\mathrm{T}}$ is a vector whose entries correspond to the N_p peaks of the mixture GC; vector $\boldsymbol{\varepsilon} = \left(\varepsilon_1, \cdots, \varepsilon_{N_p}\right)^{\mathrm{T}}$ embeds GC peak measurement errors as well as model errors; and $\underline{\underline{\mathbf{A}}}$ is a $N_p \times K$ matrix, whose entry $a_{n,k}$ is the nth peak of the kth EM of the mixture (i.e., column k of matrix $\underline{\underline{\mathbf{A}}}$ contains the N_p peaks detected for the kth EM).

Production allocation methods relying on ratios, R_{ij}, are based on the following formulation [28]

$$R_{ij} = \frac{b_i}{b_j}; \qquad \text{with} \qquad b_n = \sum_{k=1}^{K} x_k a_{n,k} \frac{m_K}{m_k}; \qquad n = i, j. \tag{2}$$

Here, m_k is the mass of the kth EM injected into the GC device. Note that m_k and x_k are (usually) unknown model parameters, to be estimated by making use of a deconvolution method (as described in Sects 2.2 and 2.3) on the basis of GC peak or ratio data.

2.2. *Deconvolution algorithms*

2.2.1. *McCaffrey algorithm*

The deconvolution algorithm proposed by McCaffrey et al. [29] rests on the peak heights mixing model and is characterized by K unknowns (i.e., the elements of vector $\mathbf{x}$). The algorithm renders an estimate of $\mathbf{x}$, $\widehat{\mathbf{x}}_{Mc}$, as

$$\widehat{\mathbf{x}}_{Mc} = \left(\underline{\underline{\mathbf{A}}}^{\mathrm{T}} \underline{\underline{\mathbf{A}}}\right)^{-1} \underline{\underline{\mathbf{A}}}^{\mathrm{T}} \mathbf{b}. \tag{3}$$

Eq. (3) descends from minimization of the generalized Least Squares (LS) criterion, assuming that elements of $\boldsymbol{\varepsilon}$ in (1) are zero-mean Gaussian random variables. To improve the accuracy of estimates based on Eq. (3), McCaffrey et al. [29] propose the following procedure:

- Evaluate (3) normalizing $a_{n,k}$ and b_n by $\max\{a_{n,1}, \ldots, a_{n,K}\}$; this enables one to properly consider the information content embedded in each peak value (even for small values of b_n).
- Determine $\boldsymbol{\varepsilon}$ by cross-validation and verify that its entries are characterized by a zero-mean Gaussian distribution and reject from the analysis peaks where entries of $\boldsymbol{\varepsilon}$ do not satisfy this condition.

Note that the McCaffrey's deconvolution algorithm corresponds to a least-squares estimation approach. The latter is widely used in several areas such as, e.g., machine learning [35], genomics [36], econometrics [37], as well as petroleum engineering (e.g., [38,39,40]). The main advantage of the algorithm is its conceptual simplicity. However, it might yield unphysical or inaccurate results when GCs of poor quality (in terms of measurement accuracy and reliability) are employed.

2.2.2. *Nouvelle algorithm*

Nouvelle et al. [28] introduce a deconvolution algorithm based on the peak ratios mixing model. The approach can be applied when (i) GCs of the EMs and of the mixtures to deconvolute as well as (ii) GC of at least one mixture with known EMs mass fractions (hereafter termed as *synthetic mixture*) are available. The main advantage of this algorithm is that common GC detection errors can be neglected since they can be significantly shadowed by relying on ratios. Additionally, considering that all components in a fluid sample are equally affected by improper storage, the mixture model relying on peak ratios is highly adaptable also to GC related to samples that have not been properly handled. Nevertheless, the application of the approach requires having at our disposal at least one synthetic mixture. Such a constraint is otherwise not needed by the McCaffrey deconvolution algorithm. This requirement implies, in turn, that efforts associated with laboratory analyses significantly increase, thus potentially limiting its application.

The mixing model is characterized by a total of ($2K - 1$) unknowns. These are subdivided into two groups: (i) the ratios of EM masses injected into the GC device, which form the entries of vector $\mathbf{MR} = (m_K/m_1, \cdots, m_K/m_{K-1})$; and (ii) the mass fractions of EMs in a mixture, i.e., $\mathbf{x}$. The Nouvelle algorithm is structured according to two steps: First (Step 1), $\mathbf{MR}$ is estimated by making use of available synthetic mixtures; then (Step 2), estimates of $\mathbf{x}$ are provided by relying on $\mathbf{MR}$ determined in Step 1.

Step 1 and Step 2 are performed by minimizing the function, $L(\,)$

$$L(\mathbf{MR}, \mathbf{x}|\mathbf{R}^{*}) = \sum_{\forall(i,j)} \ln\left\{1 + Q_{R_{ij}}^2 \left(R_{ij}^{*} - R_{ij}\right)^2\right\}, \quad \text{with}$$
$$Q_{R_{ij}} = \sqrt{\frac{2}{\sigma_{Rij}^2 (\eta_i + \eta_j)}}. \tag{4}$$

Here, $\mathbf{R}^{*}$ is a vector of components R_{ij}^{*}, the latter corresponding to the experimentally observed ratio values, R_{ij}, defined in Eq. (2); and η_i and η_j are the number of times that peaks i and j are used in the set of ratios of N_R elements, respectively. The quantity $\sigma_{R_{ij}}^2$ is the variance of R_{ij}, which is approximated in Nouvelle et al. [28] as

$$\sigma_{Rij}^2 = \frac{1}{\overline{b}_j^2} \left(\sigma_i^2 + \frac{\overline{b}_i^2}{\overline{b}_j^2} \sigma_j^2\right), \tag{5}$$

$\overline{b}_n$ and σ_n^2 (with $n = i, j$) being mean and variance of peak n, respectively. According to the procedure highlighted by Nouvelle et al. [28], $\overline{b}_n$ and σ_n^2 can be estimated on the basis of replicates of laboratory experiments. However, the number of available replicates, N, is usually very limited (typically, $N = 3$–5). This renders the accuracy and reliability of statistical moments evaluated in such a small ensemble highly questionable (note that the error associated with estimates of mean and variance decreases as N^{-1} and $N^{-0.5}$, respectively).

Nouvelle et al. [28] derived Eq. (4) by (i) assuming that measurement errors of peaks i and j can be described through a zero-mean Gaussian distribution, so that R_{ij} follows a Cauchy distribution, and (ii) determining the weighting factor $Q_{R_{ij}}^2$ relying on an approximation of $\sigma_{R_{ij}}^2$ as given by Eq. (5) instead of considering the scale parameter of the probability density function of R_{ij}. Note also that Eq. (5) relies on a Taylor expansion of Eq. (2) truncated at first order. Therefore, it is a good estimate of the variance of R_{ij} only if σ_i^2 and σ_j^2 are small.

Here, we reframe the work of Nouvelle et al. [28] within a rigorous Maximum Likelihood, ML, approach. Assuming that R_{ij} follows a Cauchy distribution, ML estimates of $\mathbf{MR}$ and $\mathbf{x}$ are obtained by minimizing the negative log-likelihood function, $NLL(\,)$

$$NLL(\mathbf{MR},\mathbf{x}|\mathbf{R}^{*}) = N_R \ln\pi + \sum_{\forall(i,j)} \ln\left\{1+\sigma_{ij}^2\left(R_{ij}^{*}-R_{ij}\right)^2\right\} - \ln\sigma_{ij} \qquad \text{with}$$

$$\sigma_{ij} = \frac{\sigma_j}{\sigma_i}. \tag{6}$$

Note that key differences between Eqs. (4) and (6) are (i) the weight factor, i.e., $Q_{R_{ij}}^2$ in Eq. (4) and σ_{ij}^2 in Eq. (6), and (ii) the additional term, $\ln\sigma_{ij}$, in (6). In the following we assume that $c_v = \sigma_n/b_n^{*}$ (b_n^{*} corresponding to the experimental value of peak n) is constant, thus implying that the relative error across peak height measurements is constant. This assumption is consistent with previous studies (e.g., [41]) linking contaminant concentration errors to measured concentration values. Thus, Eq. (6) becomes

$$NLL(\mathbf{MR},\mathbf{x}|\mathbf{R}^{*}) = N_R\ln\pi + \sum_{\forall(i,j)} \ln R_{ij}^{*} + \ln\left\{1+\left(1-\frac{R_{ij}}{R_{ij}^{*}}\right)^2\right\}. \tag{7}$$

Note that minimization of Eq. (7) is equivalent to the minimization of its last term. We further note that another possible approach is to consider σ_n as constant, i.e., independent of peak measurements. Results obtained in this case were unsuccessful and are not reported in Sect. 3.

2.3. Original approach and deconvolution algorithm

In this Section, we introduce a novel deconvolution algorithm (hereafter termed PGM, after the initials of the authors' institutions) that enables one to overcome limitations associated with the approaches described above while maintaining operational simplicity. Our approach (i) allows the use of the key concept of the ratios mixture model, i.e., explicitly considering the objective function to be based on the difference between observed and numerically evaluated ratios; (ii) does not rely on synthetic mixtures (as otherwise required by the Nouvelle deconvolution method), thus avoiding an increase in laboratory time (as compared against the McCaffrey algorithm); (iii) is characterized by theoretical foundations that enable one to overcome the assumptions and limitations required by the Nouvelle approach (as detailed in Sect. 2.2.2), and (iv) does not strictly require (in principle) replicates to obtain estimates of peak measurements variance. Our approach is conducive to estimating **x** and **MR** from the mixing model as detailed in the following.

We write

$$b_n = b_n^{*} + \lambda_n; \qquad \text{with} \qquad n = i,j, \tag{8}$$

where the peak height b_n of a mixture GC is expressed as the sum of the observed value, b_n^{*}, and a zero-mean measurement error, λ_n, characterized by a Gaussian distribution with (generally unknown) variance σ_n^2. Assuming that peak measurement errors are not correlated, R_{ij} in model is a random variable characterized by the following probability density function, pdf,

$$p_{R_{ij}}(r_{ij}) = \frac{1}{2\pi\sigma_i\sigma_j}\int_{-\infty}^{\infty}|b_j|\, e^{-\frac{1}{2}\left(\frac{r_{ij}b_j-b_i^{*}}{\sigma_i}\right)^2} e^{-\frac{1}{2}\left(\frac{b_j-b_j^{*}}{\sigma_j}\right)^2} db_j = \frac{\sigma_{ij} e^{-\frac{1}{2}\frac{b_j^{*2}}{\sigma_j^2}\left(R_{ij}^{*2}\sigma_{ij}^2+1\right)}}{\pi\left(1+r_{ij}^2\sigma_{ij}^2\right)}\left(1+\gamma_{ij}\right) \tag{9}$$

with

$$\gamma_{ij} = \sqrt{\pi}\phi_{ij}\, e^{\phi_{ij}^2} erf\,\phi_{ij};$$
$$\phi_{ij} = \frac{b_j^{*}}{\sigma_j}\frac{1+r_{ij}R_{ij}^{*}\sigma_{ij}^2}{\sqrt{2\left(1+r_{ij}^2\sigma_{ij}^2\right)}}; \tag{10}$$
$$R_{ij}^{*} = \frac{b_i^{*}}{b_j^{*}}.$$

Considering all available N_R ratios, ML estimates of model parameters (i.e., **x**, **MR**, and σ_n^2) can be obtained by minimizing the negative Log-Likelihood criterion, i.e.,

$$\begin{aligned} NLL\left(\mathbf{MR},\mathbf{x},\sigma_n^2|\mathbf{R}^{*}\right) &= -\ln\left\{p_{R_{1,2}\ldots R_{ij}\ldots}\left(r_{12},\cdots,r_{ij}\ldots|\mathbf{x},\ \mathbf{MR},\ \sigma_1,\sigma_2,\ldots.,\ \right)\right\} \\ &= -\sum_{\forall(i,j)}\ln p_{R_{ij}}\left(r_{ij}|\mathbf{x},\ \mathbf{MR},\ \sigma_i,\sigma_j\right) \end{aligned} \tag{11}$$

Note that the sum in Eq. (11) considers all of the N_R ratios in the set. Making use of Eq. (9), Eq. (11) becomes

$$NLL = N_R\ln\pi + \sum_{\forall(i,j)}\ln\left(\frac{1+r_{ij}^2\sigma_{ij}^2}{\sigma_{ij}}\right) + \frac{1}{2}\frac{b_j^{*2}}{\sigma_j^2}\left(R_{ij}^{*2}\sigma_{ij}^2+1\right) - \ln\left(1+\gamma_{ij}\right). \tag{12}$$

As in Sect. 2.2.2, we assume that $c_v = \sigma_n/b_n^{*}$ is constant across peak height measurements and Eq. (12) simplifies as

$$NLL = J + N_R\left(\ln\pi + \frac{1}{c_v^2}\right); \quad \text{with} \quad J = \sum_{\forall(i,j)}\ln R_{ij}^{*} + \ln\left(1+\frac{r_{ij}^2}{R_{ij}^{*2}}\right) - \ln\left(1+\upsilon_{ij}\right), \tag{13}$$

where

$$\upsilon_{ij} = \sqrt{\pi}\omega_{ij}\, e^{\omega_{ij}^2} erf\,\omega_{ij};$$
$$\omega_{ij} = \frac{1+r_{ij}/R_{ij}^{*}}{c_v\sqrt{2\left(1+r_{ij}^2/R_{ij}^{*2}\right)}}. \tag{14}$$

Parameters embedded in Eq. (13) include **x**, **MR**, and c_v. If c_v is known, minimization of *NLL* (Eq. (13)) coincides with minimization of *J*. If c_v is unknown, its ML estimate can be obtained according to

$$\frac{\partial NLL}{\partial c_v} = -\frac{2N_R}{c_v^3} + \frac{1}{c_v}\sum_{\forall(i,j)}\frac{1}{\left(1+\upsilon_{ij}\right)}\left(2\omega_{ij}^2 + 2\omega_{ij}^2\upsilon_{ij} + \upsilon_{ij}\right) = 0. \tag{15}$$

One can then evaluate c_v by solving the following implicit equation

$$c_v^{-2} = \frac{1}{N_R}\sum_{\forall(i,j)}\omega_{ij}^2 + \frac{\upsilon_{ij}}{2\left(1+\upsilon_{ij}\right)}. \tag{16}$$

Here, we propose to obtain ML estimates of **x**, **MR**, and c_v (denoted as $\widehat{\mathbf{x}}_{PGM}$, $\widehat{\mathbf{MR}}_{PGM}$ and $\widehat{c}_{v,PGM}$, respectively) according to the procedure highlighted in the flowchart depicted in Fig. 1. The latter shows that the procedure requires to (i) initialize **x** and **MR**; (ii) compute the ratios of the mixture chromatogram using Eq (2); (iii) initialize the value of c_v and minimize J in Eq. (13) to compute the ML estimates $\widehat{\mathbf{x}}_{PGM}$ and $\widehat{\mathbf{MR}}_{PGM}$; and (iv) make use of Eq. (16) to evaluate $\widehat{c}_{v,PGM}$. The procedure ends when a convergence criterion (e.g., $\left|c_v-\widehat{c}_{v,PGM}\right|/c_v < \delta_0$) is satisfied. Note that δ_0 is a threshold value that must be defined at the beginning of the workflow. In our test case, we set $\delta_0 = 0.01$. Moreover, we note that in our analyses the same mass of EMs samples is employed during GC experiments. As such, the ratios of EM masses injected into the chromatography device (corresponding to the entries of vector **MR**) are constrained to the range 0.95–1.05 during the optimization procedure.

2.4. Alternating least squares algorithm, ALS

All deconvolution methods described in Sects. 2.2 and 2.3 allow estimating the mass fraction of EMs in a mixture when chromatograms of EMs (i.e., entries of matrix $\underline{\mathbf{A}}$) are known. The ALS deconvolution algorithm tackles a fundamentally different production allocation problem in which entries of matrix $\underline{\mathbf{A}}$ are not known. Nevertheless, the application of ALS requires the analysis of multiple virtual mixtures (the number of which must be greater than the number of EMs) that need to

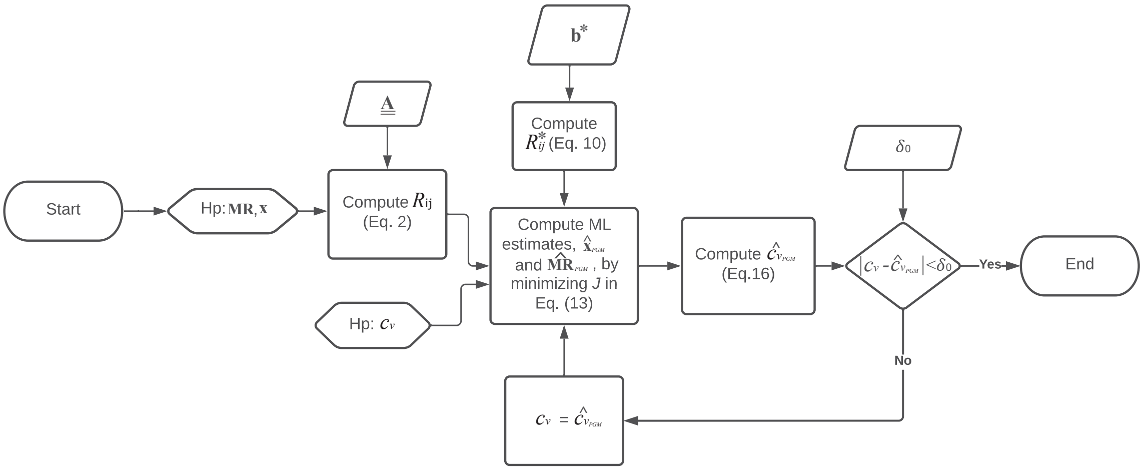
Start
Hp: MR, x
A
Compute R_{ij} (Eq. 2)
Hp: c_v
b*
Compute R^*_{ij} (Eq. 10)
Compute ML estimates, $\hat{x}_{PGM}$ and $\widehat{MR}_{PGM}$, by minimizing J in Eq. (13)
Compute $\hat{c}_{v_{PGM}}$ (Eq.16)
δ_0
$|c_v - \hat{c}_{v_{PGM}}| < \delta_0$
Yes
End
No
$c_v = \hat{c}_{v_{PGM}}$

Fig. 1. Flowchart of the PGM deconvolution algorithm.

be subject to the deconvolution simultaneously. To provide a proper representation of the overall system variability, these virtual mixtures must be associated with different mass fraction compositions of EMs.

As ALS relies on multiple mixtures, vectors **x** and **b** in model are now matrices. These are hereafter denoted as $\underline{\underline{\mathbf{x}}}$ and $\underline{\underline{\mathbf{b}}}$ of size ($K \times N_M$) and ($N_p \times N_M$), respectively (N_M being the number of mixtures being deconvoluted simultaneously).

The first step in an ALS-based production allocation relies on estimating the number of EMs, K, making use of one of the following methodologies (or a combination thereof):

(i) perform a principal component analysis, PCA (or a Singular Value Decomposition – SVD), of the mixture GCs included in $\underline{\underline{\mathbf{b}}}$ and evaluate the minimum number of components required to explain the variance of $\underline{\underline{\mathbf{b}}}$ [42,43];

(ii) make use of information about the natural system (e.g., the geological structure) that could assist in constraining K.

It is noted that, when the assessment of K is not unambiguous, the deconvolution process should be performed several times, upon varying K and analyzing the ability of the estimated EM spectra and mass fractions to describe $\underline{\underline{\mathbf{b}}}$ [42].

Once K is defined, the ALS deconvolution algorithm rests on an iterative procedure to determine a Least Square (LS) estimate of $\underline{\underline{\mathbf{A}}}$ and $\underline{\underline{\mathbf{x}}}$ according to the following workflow:

Step 1. Initialize $\underline{\underline{\mathbf{A}}}$.

Step 2. Determine an LS estimate of $\underline{\underline{\mathbf{x}}}$ as

$$\hat{\underline{\underline{\mathbf{x}}}}_{ALS} = \left(\underline{\underline{\mathbf{A}}}^{\mathrm{T}}\underline{\underline{\mathbf{A}}}\right)^{-1}\underline{\underline{\mathbf{A}}}^{\mathrm{T}}\underline{\underline{\mathbf{b}}}. \tag{17}$$

Step 3. Determine an LS estimate of $\underline{\underline{\mathbf{A}}}$ as

$$\hat{\underline{\underline{\mathbf{A}}}}_{ALS} = \underline{\underline{\mathbf{b}}}\,\hat{\underline{\underline{\mathbf{x}}}}^{T}_{ALS}\left(\hat{\underline{\underline{\mathbf{x}}}}_{ALS}\ \hat{\underline{\underline{\mathbf{x}}}}^{T}_{ALS}\right)^{-1}, \text{ with entries } (a_{ALS})_{n,k} > 0. \tag{18}$$

Step 4. If $\left|a_{n,k} - (a_{ALS})_{n,k}\right| \big/ a_{n,k} < \delta_0$, then stop; otherwise, set $\underline{\underline{\mathbf{A}}} = \hat{\underline{\underline{\mathbf{A}}}}_{ALS}$ and go to step 2.

Note that, since GCs with negative entries have no physical meaning, entries of $\hat{\underline{\underline{\mathbf{A}}}}_{ALS}$ (Eq. (18)) are constrained to be positive. Note also that N_M must be larger than (or equal to) K, to guarantee that the system is not under-constrained.

The workflow described above must be repeated multiple times with different initializations of $\underline{\underline{\mathbf{A}}}$ (Step 1) to avoid entrapment in local minima of the objective function to be minimized [44]. We note that, since Eq. (17) for the evaluation of $\hat{\underline{\underline{\mathbf{x}}}}_{ALS}$ is coupled with, it is possible that the global minimum value of the objective function is not necessarily associated with the optimal $\underline{\underline{\mathbf{x}}}$ (i.e., associated with the minimum LS value) due to the action of $\hat{\underline{\underline{\mathbf{A}}}}_{ALS}$ in. To overcome this issue, we propose an original view and frame the approach within a probabilistic Monte Carlo setting. We do so by relying on multiple realizations. Each one of these corresponds to a combination of the N replicates associated with GC performed for each mixture (to be deconvoluted simultaneously). To the best of our knowledge, this is the first study exploring the potential of the ALS deconvolution algorithm in the context of a production allocation scenario under such a probabilistic framework.

2.5. Experimental setup

The analysis of the relative skills of the deconvolution algorithms described in Sects. 2.2–2.4 to assist production allocation is assessed upon considering a set of ten laboratory-based mixtures produced by commingling three EMs as listed in Table 1. These mixtures have been selected with the aim of reproducing at the laboratory scale typical commingling scenarios associated with field settings.

2.5.1. Materials

All solvents, including dichloromethane, are of analytical grade. These, as well as ethylbenzene-d10 (used as an internal standard for quantification purposes), were purchased from Merck (Darmstadt, Germany). An alkylbenzene standard mixture containing 37 compounds and used for identification, method development, and evaluation of response factors associated with the internal standard was purchased from Restek Corporation (Bellefonte, United States). Oil mixtures and EMs were properly stored in a fridge at a temperature of 4 °C to minimize potential alterations due to improper handling of the samples.

Table 1
Mass fractions of the three EMs in the 10 laboratory-based mixtures.

Oil sample	EM1	EM2	EM3
	x_1 [%]	x_2 [%]	x_3 [%]
M1	33.3	33.3	33.3
M2	70	15	15
M3	10	70	20
M4	20	20	60
M5	50	30	20
M6	20	40	40
M7	45	55	0
M8	5	10	85
M9	85	10	5
M10	0	90	10

2.5.2. Sample preparation and GC–MS analysis

EM samples are weighed and dissolved in dichloromethane to a concentration of 10 mg/mL and used to produce the 10 laboratory mixtures illustrated in Table 1. All EMs samples and mixtures are then analyzed 5 times through GC–MS by targeting the alkylbenzene components in the molecular range C_8-C_{12}. Fixed amounts of ethylbenzene-d10 are added to each sample to assist quantification of the various alkylbenzenes. This step is performed by applying the internal standard method, using peak heights and response factors evaluated from a standard alkylbenzene mixture of 37 compounds. Since oil samples contain a number of alkylbenzenes that is significantly higher than what is available in the standard mixture, response factors equal to those of the most closely eluting alkylbenzene available as standard are assigned to such compounds.

The analysis of alkylbenzene compounds in the oil samples is performed via gas chromatography-single quadrupole mass spectrometry (GC–MS). Helium is used as carrier gas and the injections are performed in split mode. The analytical separation is carried out using a Stabilwax capillary column (Restek – 60 m × 0.32 mm × 0.25 μm) in temperature gradient mode. The eluted compounds are ionized within the electron ionization source of the mass spectrometer, which operates at 70 eV and 250 °C source temperature.

The MS analyzer is operated in full scan mode only in the early stages of method development. This yields spectral data for the identification and quantification of alkylbenzene components in the standard mixture. Otherwise, sample analyses are conducted in SIM (Selected Ion Monitoring) mode. Mass-to-charge ratios (m/z) for SIM acquisition are: 91, 105, 106, 116, 119, 120, 133, 134, 147, 148, 162. A quantitative analysis is performed for each peak using the measured heights in the GC.

3. Results

3.1. Experimental method development and available data

As discussed in Sect. 1, several methods have been reported in the literature, including HRGC, MDGC, and GC–MS analysis. The current analyses are performed using GC–MS since (i) GC–MS requires a simpler and more accessible instrumentation and allows for a more selective analysis of alkylbenzenes than MDGC, especially if high molecular weight compounds need to be assessed; (ii) GC–MS analyses are faster than their counterparts based on MDGC or HRGC; (iii) selectivity of GC–MS allows filtering out all signals related to non-alkylbenzene species which are typically observed in oil samples, replacing the need for a double separative column (which is otherwise required for MDGC); (iv) GC–MS allows monitoring additional compound classes with respect to MDGC (e.g., diamondoids, alkylnaphtalenes, dibenzothiophenes, polycyclic aromatic hydrocarbons) further increasing the number of geochemical parameters that can be obtained for fingerprinting or other applications; and (v) GC–MS yields a complete baseline separation of the 11 alkylbenzenes analyzed by MDGC and extends the analytical range up to C_{12}-alkylbenzenes, thus enabling one to analyze 50–80 compounds. In this context, advantages related to our operational choice include: (i) the possibility to identify more discriminating features in case of highly similar oils, with an ensuing increase in the quality of production allocation estimates; and (ii) the observation that the higher boiling point molecules better compensate for poor sampling practices or improper storage conditions.

It is otherwise noted that extending the analytical range beyond the C_9-alkylbenzenes has the potential drawback of increasing the number of isomers and the possibility of coelutions. This potentially renders the chromatographic separation more complex and the chromatographic peaks less resolved. A solution to this issue can be found upon relying on the selectivity attributes of mass spectrometry, which allows for easier discrimination. For example, doing so enables one to separate C_{12}-alkylbenzenes from C_{11}-alkylbenzenes due to their associated differing m/z ratios (133 vs 119, respectively). Therefore, even as a complete GC separation is virtually impossible, a sufficiently high number of well-resolved peaks can be readily identified. This in turn allows enhancing the discriminatory capacity of the deconvolution algorithms employed for production allocation estimates.

Analysis and comparison of several oil samples associated with various sources suggest that the analyzed alkylbenzenes do not suffer from contamination or interferences (details not shown). Therefore, the methodology does not require a manual peak selection, because it always targets the same suite of compounds. Thus, these are automatically quantified and fed to the deconvolution algorithms.

Measurement precision plays a key role in GC fingerprinting production allocation studies, as a successful production allocation is closely linked to the ability to effectively distinguish EMs from the samples analyzed. The use of compounds (alkylbenzenes) associated with almost identical chemical and physical properties further enhances instrumental precision. In our study, the coefficient of variation of component measurements is generally lower than 5% and never exceeds 10%. As an example, Fig. 2 depicts the chromatogram of the five replicates associated with mixture M1. These experimental results suggest an overall high degree of repeatability of the experimental analyses. The remaining mixtures and EMs display a similar quality of repeatability (details not shown). One can see that peak responses vary across two orders of magnitude, the largest values being associated with the first 11 peaks (i.e., those related to C_9-alkylbenzenes).

3.2. Deconvolution

Here, we present a quantitative comparison of the accuracy of the various deconvolution algorithms illustrated in Sect. 2 on the basis of the laboratory dataset detailed in Sect. 3.1. For this purpose, we compute estimates of $\mathbf{x}$, $\widehat{\mathbf{x}}_\xi$, using (i) Eq. (3) when $\xi = Mc$ (McCaffrey algorithm), (ii) Eq. (4) when $\xi = Nv$ (Nouvelle algorithm), (iii) Eq. (7) when $\xi = MNv$ (modified Nouvelle algorithm), (iv) Eqs. (13)–(16) when $\xi = PGM$, and (v) Eqs. (17)–(18) when $\xi = ALS$. For the McCaffrey and PGM algorithms no synthetic mixtures are required. Thus, estimates $\widehat{\mathbf{x}}_\xi$ for each mixture have been evaluated considering all combinations of the N replicates of the given mixture and associated EMs chromatograms. This yields N^{K+1} values of $\widehat{\mathbf{x}}_\xi$. On the other hand, for the ALS algorithm (where N_M is required to be larger than K) a subset of the N^{N_M} possible estimates has been randomly selected. With reference to the Nouvelle algorithm (where at least one synthetic mixture is required), we explore the goodness of the deconvolution algorithm by varying the number of synthetic mixtures, N_{SM}, from 1 to 9.

The mean associated with estimates $\widehat{\mathbf{x}}_\xi$ (i.e., $\overline{\mathbf{x}}_\xi$) is then evaluated upon considering that the available data are compositional vectors (whose components express proportions or percent amount of a whole). We then leverage on the theoretical framework underlying Compositional Data Analysis (CoDa; e.g., [45–47] and references therein). For comparison purposes, we also provide an estimate of $\mathbf{x}_\xi$ (denoted $\overline{\overline{\mathbf{x}}}_\xi$) by averaging EMs and mixture replicates before implementing a deconvolution algorithm.

Finally, we assess the performance of each deconvolution method by computing the mean absolute error, MAE_ξ, and the mean absolute percentage error, $MAPE_\xi$, obtained with deconvolution method ξ and defined as

$$MAE_\xi = \frac{1}{K}\sum_{k=1}^{K}\left|x_k^* - \overline{\overline{x}}_{k,\xi}\right|; \qquad MAPE_\xi = \frac{100}{K}\sum_{\substack{k=1 \\ x_k \neq 0}}^{K}\left|\frac{x_k^* - \overline{\overline{x}}_{k,\xi}}{x_k^*}\right|. \tag{19}$$

Here, x_k^* is the true value associated with the kth EM in a mixture and $\overline{\overline{x}}_{k,\xi}$ is the kth element of $\overline{\overline{\mathbf{x}}}_\xi$. Note that by making use of $\overline{\mathbf{x}}_\xi$ instead of $\overline{\overline{\mathbf{x}}}_\xi$ in (19) we obtained almost identical results (details not shown).

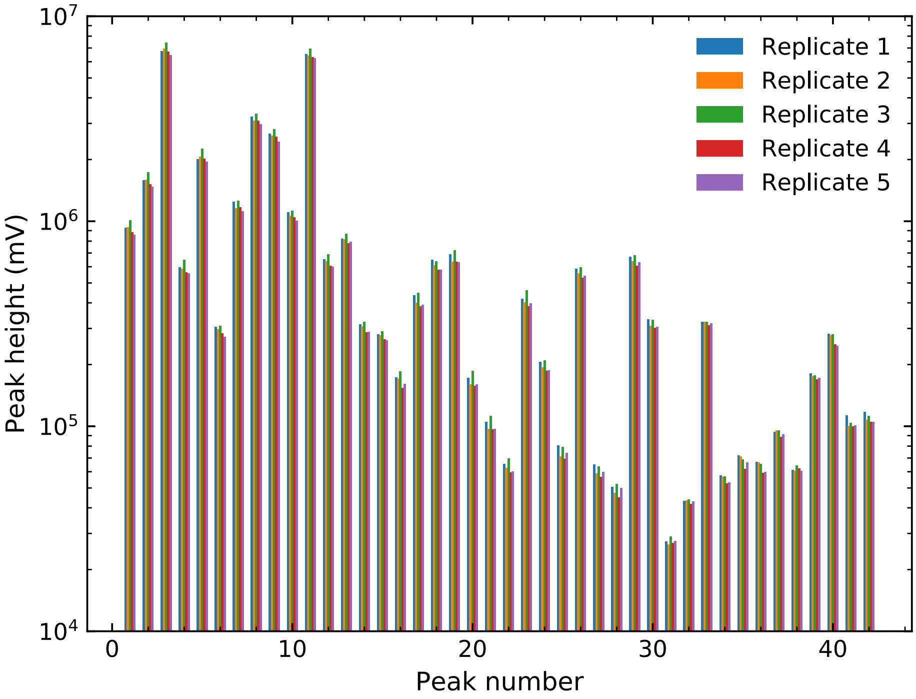

Fig. 2. Chromatograms resulting from 5 replicates (Mixture M1).

3.2.1. *McCaffrey deconvolution method*

Fig. 3 provides ternary diagrams of the N^{K+1} estimates $\widehat{\mathbf{x}}_{Mc}$ (gray symbols). The variability of $\widehat{\mathbf{x}}_{Mc}$ is modest, samples M7, M9, and M10 being the only exceptions. Note that M7 and M10 are formed by only two EMs (i.e., $K = 2$). The compositional mean associated with these estimates, $\overline{\mathbf{x}}_{Mc}$ (black cross), experimental values, $\mathbf{x}^*$ (green circle), as well as values of the deconvolution obtained by averaging EMs and mixture GC replicates before the use of the deconvolution algorithm, $\overline{\overline{\mathbf{x}}}_{Mc}$, are also included in Fig. 3. Quantities $\overline{\mathbf{x}}_{Mc}$ and $\overline{\overline{\mathbf{x}}}_{Mc}$ are seen to properly represent the overall behavior of the EMs mass fractions for all of the mixtures tested. Across all mixtures, the average MAE_{Mc} (Eq. (19)) is 4.3% (its corresponding median being equal to 3%), and the average $MAPE_{Mc}$ is equal to 22.9% (median being equal to 11.9%) . Fig. 4 depicts histograms of Aitchison distances between measured mass fractions $\mathbf{x}^*$

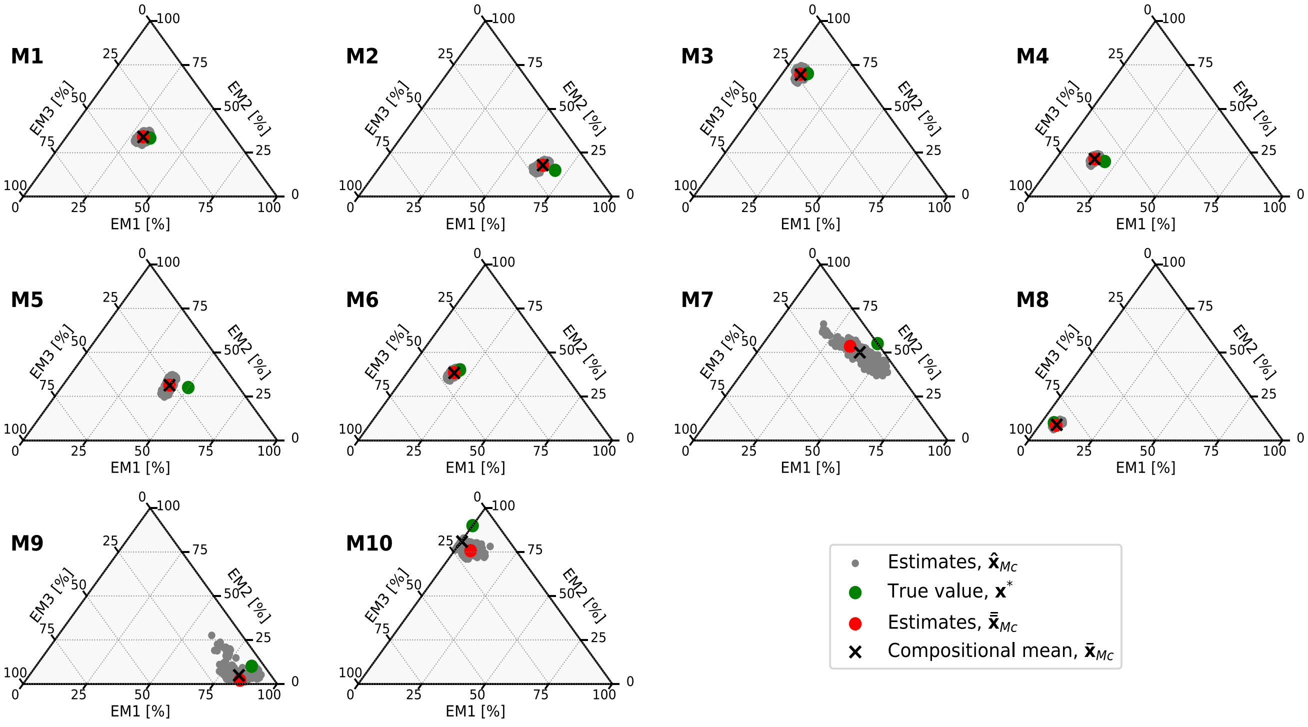

Fig. 3. Results of the production allocation approach obtained through the McCaffrey deconvolution algorithm. Each ternary diagram corresponds to a given mixture and includes: (i) Individual Estimates, $\widehat{\mathbf{x}}_{Mc}$; (ii) Compositional mean of $\widehat{\mathbf{x}}_{Mc}$, $\overline{\mathbf{x}}_{Mc}$; (iii) The true value, $\mathbf{x}^*$; and (iv) Estimates obtained by averaging mixtures and EMs replicates before performing the deconvolution, $\overline{\overline{\mathbf{x}}}_{Mc}$.

and their estimated counterparts $\widehat{\mathbf{x}}_{Mc}$. These results confirm quantitatively the observations made to Fig. 3 according to which the largest variabilities of the estimates are associated with mixtures M7, M9, and M10, these mixtures being also associated with the largest mean estimation errors.

3.2.2. Nouvelle deconvolution approach

Here, we analyze four test cases (C1, C2, C3, and C4; see Table 2). Test cases C1 and C3 consider ratios of two consecutive peaks (e.g., b_1/b_2,b_2/b_3,...) and two formulations of the objective function employed in the procedure, corresponding to the Nouvelle et al. [28] deconvolution method (Eq. (4)) and our suggested modification (Eq. (7)), respectively. Then, in test cases C2 and C4 we explore the benefit of relying on ratios evaluated upon using more than two peaks (e.g., $b_1/(b_2+b_3)$) for the deconvolution procedure.

Fig. 5 depicts box plots of $MAPE_{Nv}$ values for mixture M3 (the remaining tested mixtures exhibit a similar behavior; details not shown) for an increasing number of the synthetic mixtures, N_{SM}, used for the deconvolution algorithm. Mean and median values of $MAPE_{Nv}$ are also included. Note that results plotted in Fig. 5 include all possible combinations of synthetic mixtures (i.e., 9, 36, 84, 126, 126, 84, 36, 9, and 1 combinations of synthetic mixtures when $N_{SM}=1,2,\ldots,9$, respectively).

As expected, increasing N_{SM} tends to yield errors which are smaller and associated with reduced variability in all test cases. These errors are concentrated in the 5%-25% range for C4. In general, when a sufficient number of synthetic mixtures is available (at least 3 according to our analyses), all test cases are associated with production allocation estimates that are as accurate as those stemming from the McCaffrey deconvolution algorithm.

When the set formed by ratios between two consecutive peaks is used (i.e., scenarios C1 and C3), similar results are obtained employing the deconvolution algorithm proposed by Nouvelle et al. [28] and our proposed modification (i.e., Eq. (7)). Otherwise, if the set involving ratios of more than two peaks is considered (i.e., scenarios C2 and C4), results of enhanced accuracy are achieved upon relying on the proposed objective function given by Eq. (7) than on the original Nouvelle algorithm. Note that case C4 is characterized by the overall best accuracy of the results even when only one synthetic mixture is available. This suggests that by relying on the objective function we propose (Eq. (7)) one can potentially reduce the laboratory efforts associated with the preparation and analysis of synthetic mixtures whilst the precision of the method is enhanced. This is an important observation, as time constraints can limit the number of synthetic mixtures available in practical production allocation applications.

Table 2
Scenarios used for the assessment of the Nouvelle-based deconvolution approach.

	Ratios of two peaks	Ratios of more than two peaks
Eq. (4)	C1	C2
Eq. (7)	C3	C4

3.2.3. Original PGM approach and algorithm

Fig. 6 provides ternary diagrams of the N^{K+1} estimates $\widehat{\mathbf{x}}_{PGM}$ (gray symbols). Fig. 6 also includes the compositional mean of $\widehat{\mathbf{x}}_{PGM}$, $\overline{\mathbf{x}}_{PGM}$, experimental values, $\mathbf{x}^*$, as well as values of the deconvolution obtained by averaging EMs and mixture GC replicates before the use of the deconvolution algorithm, $\overline{\overline{\mathbf{x}}}_{PGM}$. Similar to Fig. 3, values of $\overline{\mathbf{x}}_{PGM}$ and $\overline{\overline{\mathbf{x}}}_{PGM}$ are close to $\mathbf{x}^*$. Notably, also mixtures M7, M9, and M10 are associated with small estimation errors (as opposed to what is noted in Fig. 3). Across all mixtures, the average MAE_{PGM} (Eq. (19)) is only 2.5% (its corresponding median being equal to 2.1%) and the average $MAPE_{PGM}$ is equal to 11.7% (its corresponding median being equal to 6.0%). The variability of the individual estimates is modest and significantly smaller than the one displayed in Fig. 3. Fig. 7 presents histograms of the Aitchison distances between measured mass fractions $\mathbf{x}^*$ and $\widehat{\mathbf{x}}_{PGM}$. It is noted that the mean of the Aitchison distances rendered by our original PGM deconvolution algorithm is significantly smaller than its counterpart related to the McCaffrey deconvolution algorithm for almost all mixtures analyzed. Mixtures M7, M9, and M10 are characterized by the largest values of the Aitchison distance. This is similar to what has been documented for the results of the McCaffrey deconvolution algorithm,

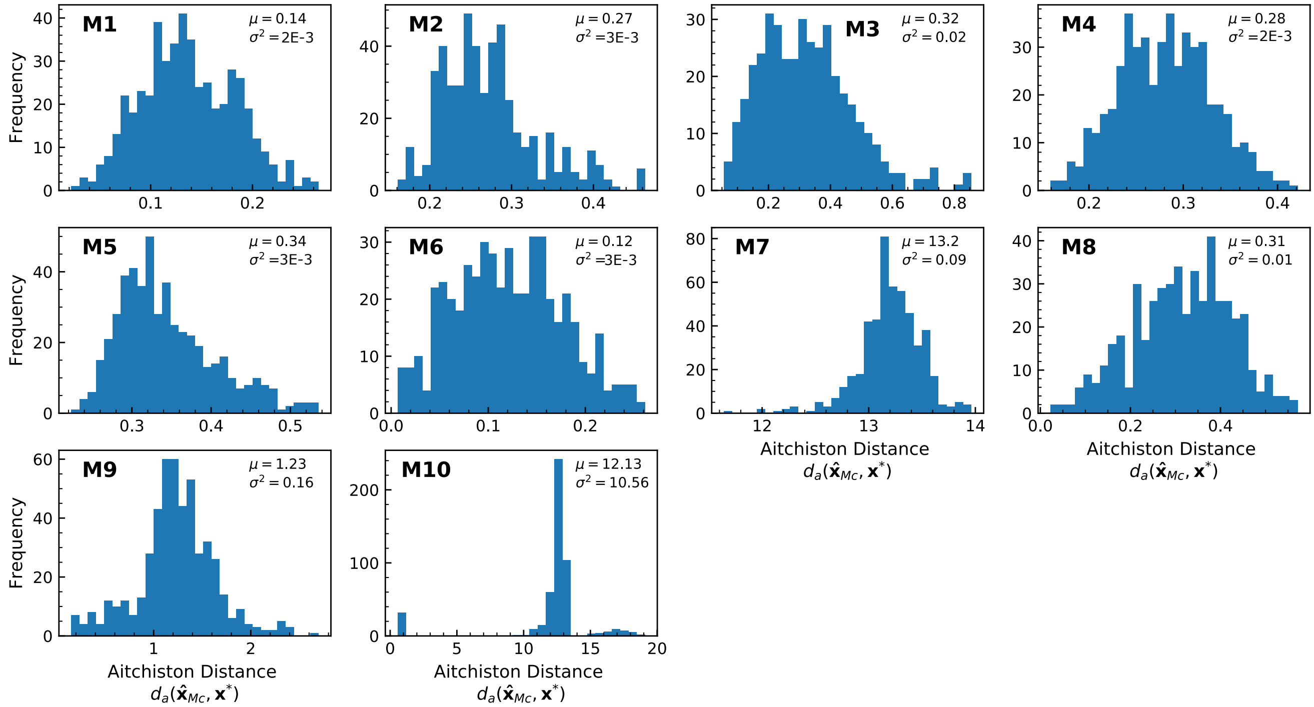

Fig. 4. Histograms of the Aitchison distances between measured mass fractions $\mathbf{x}^*$ and their estimated counterparts $\widehat{\mathbf{x}}_{Mc}$. Values of the resulting mean (μ) and variance (σ^2) are also listed.

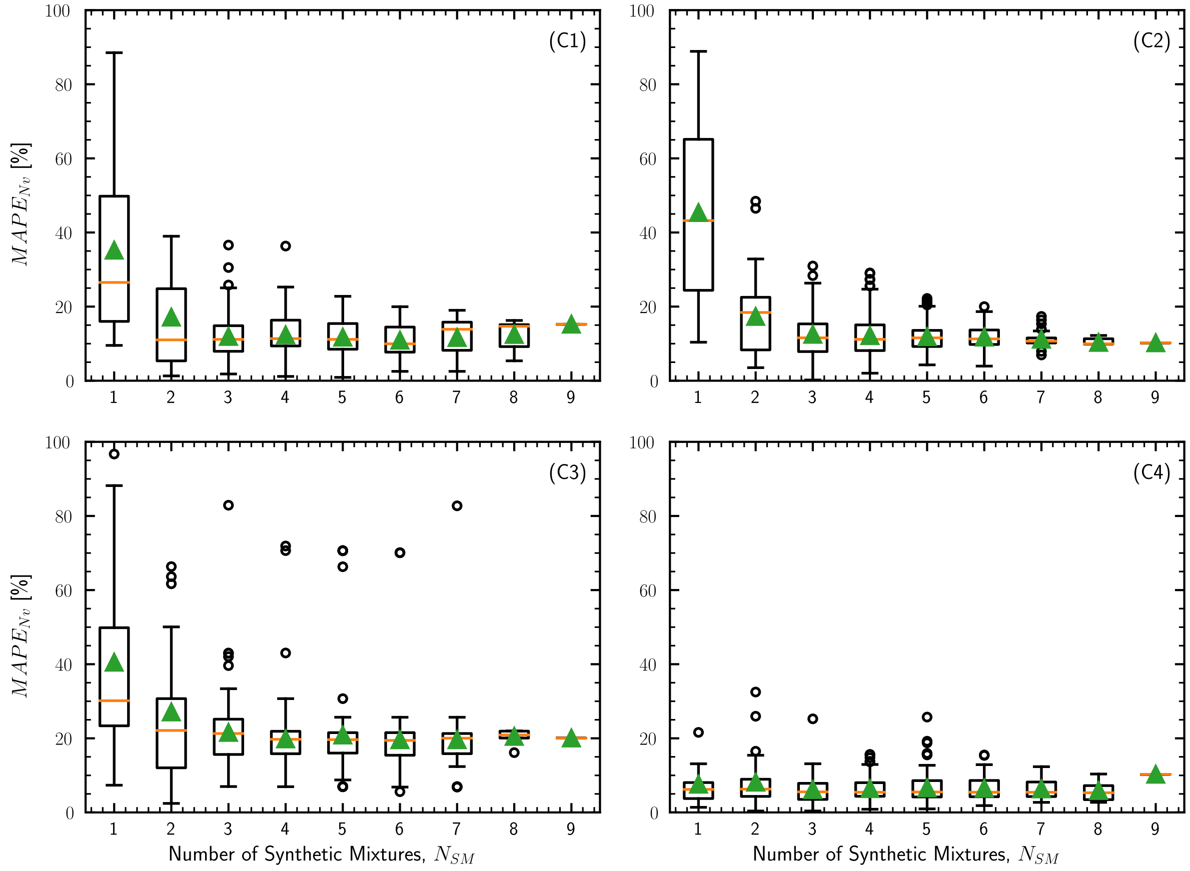

Fig. 5. Box plots of $MAPE_{Nv}$ for mixture M3 versus the number of synthetic mixtures used in the algorithm N_{SM}. Green triangles and orange lines represent the median and the mean of the distribution, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

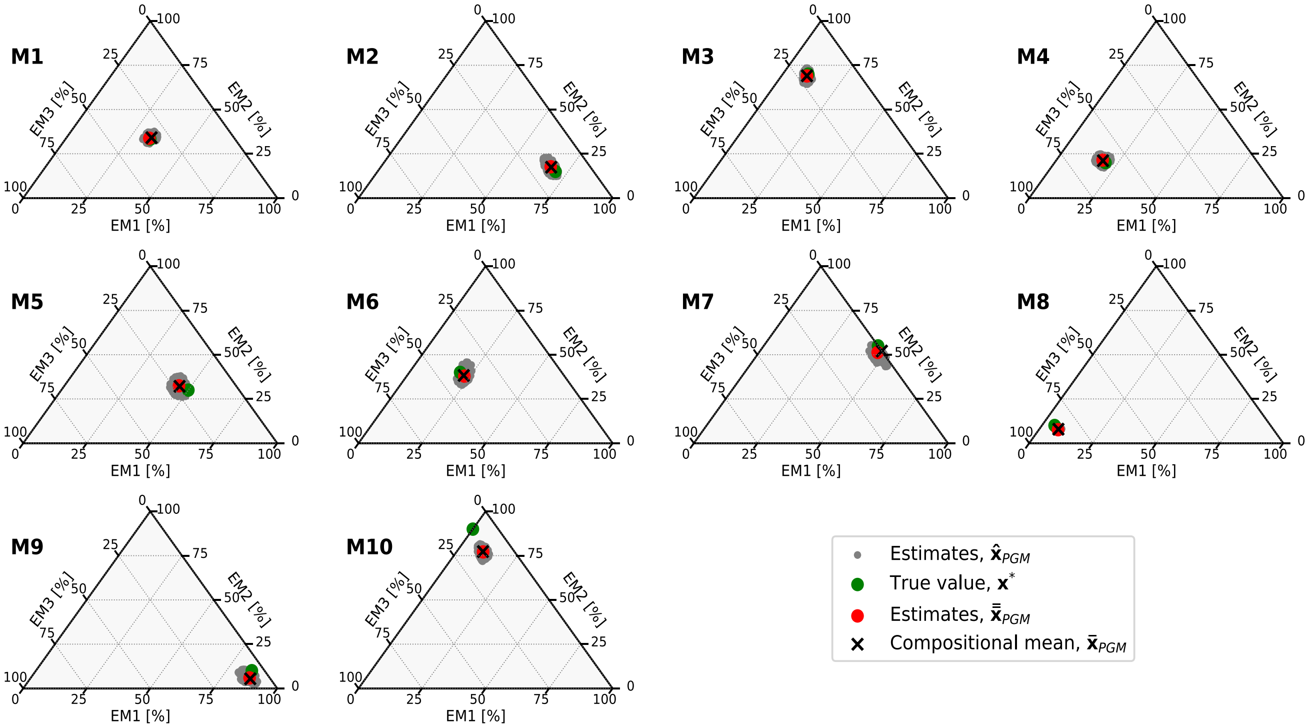

Fig. 6. Results of the production allocation approach obtained through our original PGM approach. Each ternary diagram corresponds to a given mixture and includes: (i) Individual Estimates, $\hat{\mathbf{x}}_{PGM}$, (ii) Compositional mean of $\hat{\mathbf{x}}_{PGM}$, $\bar{\mathbf{x}}_{PGM}$; (iii) True value, $\mathbf{x}^*$; and (iv) Estimates obtained by averaging mixtures and EMs replicates before performing the deconvolution, $\bar{\bar{\mathbf{x}}}_{PGM}$.

even as a reduced variability can be observed here.

3.3. Production allocation without knowledge of end members and use of the ALS algorithm

In cases where some (or all) of the chromatograms of EMs are not available and the use of the approaches and algorithms illustrated in Sects. 2.2 and 2.3 is hampered, production allocation can be performed upon relying on the ALS deconvolution algorithm as described in Sect. 2.4.

To evaluate the number of EMs, K, we perform an SVD of the mixture GCs included in $\underline{\mathbf{b}}$. Our results reveal that considering 1, 2, and 3 EMs can explain 78.9%, 96.8%, and 99.4% of the variance of $\underline{\mathbf{b}}$, respectively (details not shown). The selection of three EMs is therefore well justified by our data, which are indeed formed by two or three EMs (see Table 1).

We explore the effect of the size S of a subset of the N^{N_M} possible combinations of replicates of mixtures' GC on the stability and accuracy of the estimator $\underline{\bar{\mathbf{x}}}_{ALS}$ by plotting in Fig. 8 the quantity $\frac{1}{K}\sum_{k=1}^{K}\left|x_k^* - \bar{x}_{k,ALS}\right|$ versus S ($K = 3$, except for M7 and M10 where $K = 2$). Our results suggest that (a) the compositional mean of the estimates, $\underline{\bar{\mathbf{x}}}_{ALS}$, as well as the error between $\bar{\mathbf{x}}_{ALS}$ and $\mathbf{x}^*$ tend to stabilize by increasing the number of realizations; and (b) relying on about 1000 realizations yields stable results of the quantities of interest.

Fig. 9 depicts estimates $\underline{\widehat{\mathbf{x}}}_{ALS}$ associated with $S = 1000$ random combinations of the mixtures' GC replicates. Each combination $\underline{\mathbf{A}}$ is randomly initialized 10^3 times. The best one (i.e., the one that minimizes Eq. (17)) is selected and plotted in Fig. 9 for each combination (gray symbols). The compositional mean associated with estimates $\underline{\bar{\mathbf{x}}}_{ALS}$ (black cross), experimental values $\underline{\mathbf{x}}^*$ (green circle) as well as values of the deconvolution obtained by averaging mixtures' GC replicates before the use of the deconvolution algorithm, $\underline{\bar{\bar{\mathbf{x}}}}_{ALS}$ (red circle), are also included in Fig. 9. Values of $\underline{\bar{\mathbf{x}}}_{ALS}$ and $\underline{\bar{\bar{\mathbf{x}}}}_{ALS}$ are close to the true values of mass fractions of EMs in each mixture, although with reduced accuracy when compared against results of deconvolution algorithms based on EMs' chromatograms. The average MAE_{ALS} across mixtures (Eq. (19)) is 9.8% (its corresponding median being equal to 10.2%) and the average $MAPE_{ALS}$ is 46.4% (the median being equal to 33.8%). Fig. 10 depicts histograms of the Aitchison distances between measured mass fractions $\underline{\mathbf{x}}^*$ and rows of $\underline{\widehat{\mathbf{x}}}_{ALS}$. Our results indicate that mean Aitchison distances associated with the results rendered by the ALS algorithm are in general larger than their counterparts stemming from the McCaffrey and PGM algorithms. The largest distances are associated with mixtures M9 and M8.

3.4. Performance of the analyzed deconvolution algorithms

Table 3 lists values of the MAE_ξ and $MAPE_\xi$ metrics related to M3 and associated with production allocation estimates obtained through all deconvolution algorithms discussed in this study. The importance of the number of features (i.e., peaks) used by the deconvolution algorithms is also tested through the comparison of the resulting allocation errors by employing (i) the entire available set of 41 peaks of the mixture GCs and (ii) a subset of the first 11 peaks from the original dataset. Note that (a) the Nouvelle deconvolution algorithm (implemented by considering the original formulation, Eq. (4), or our proposed modification, Eq. (7)), is applied using M1 as synthetic mixture and employing ratios encompassing more than two peaks; and (b) the ALS deconvolution algorithm is initialized for a total of 10^3 times.

The accuracy of the estimates tends to increase with the number of features (either peak height or peak ratios) employed. Thus, the additional information content carried by considering various peaks (which might include enhanced information on molecular differences between oils or mitigate the effect of measurement errors) can be beneficial to enhance the accuracy of production allocation estimates. Therefore, extending the target alkylbenzene molecular range to C_{12}-species with the experimental procedure illustrated in this study leads to enhanced robustness and accuracy of all deconvolution methods.

In general, our proposed original approach, as well as our reformulation of the Nouvelle algorithm, are characterized by an improved performance (in terms of the metrics considered in this study) when compared against the traditionally employed deconvolution algorithms.

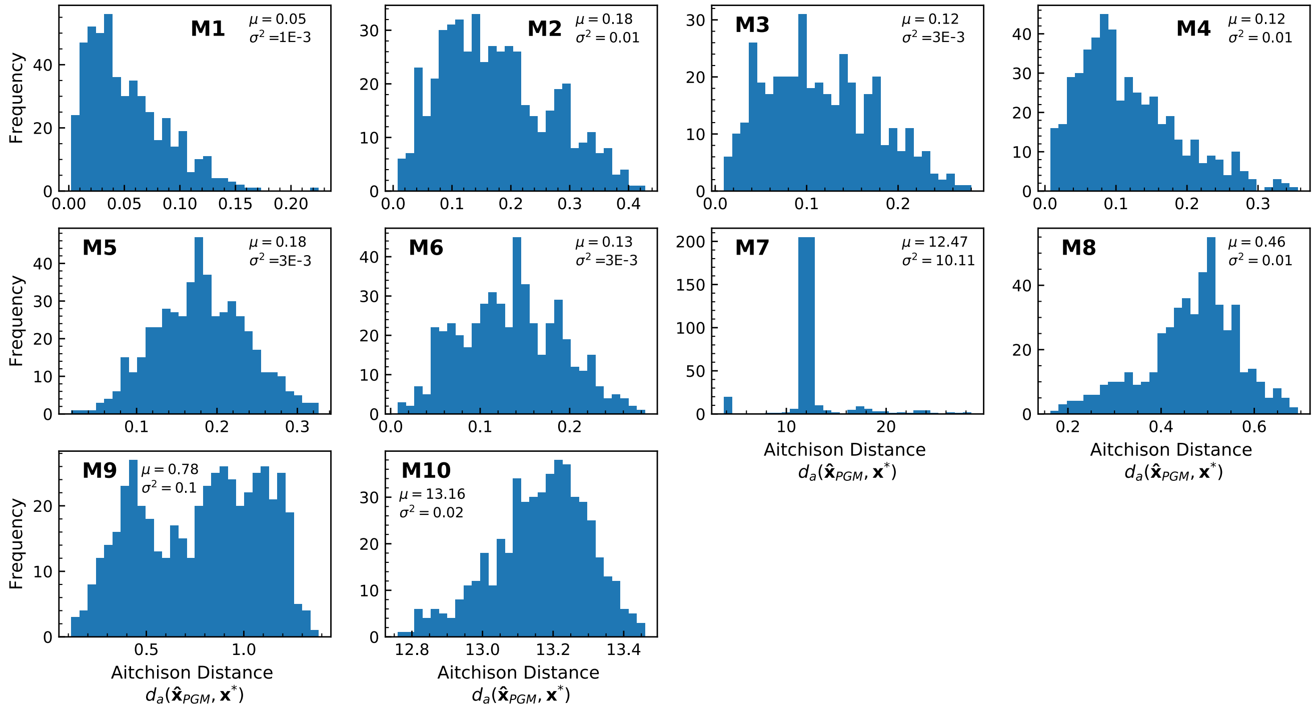

Fig. 7. Histograms of the Aitchison distances between measured mass fractions $\mathbf{x}^*$ and $\widehat{\mathbf{x}}_{PGM}$. Values of the resulting mean (μ) and variance (σ²) are listed as reference.

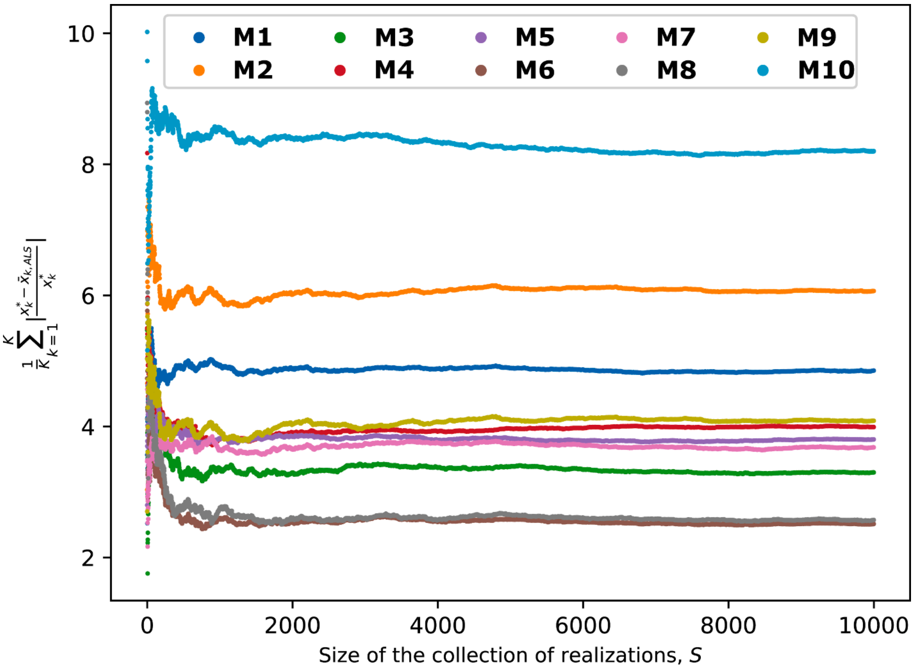

Fig. 8. Evolution of $\frac{1}{K}\sum_{k=1}^{K}\left|x_k^* - \bar{x}_{k,ALS}\right|$ for increasing size of the collection of realizations S used for the evaluation of the compositional mean in the ALS algorithm.

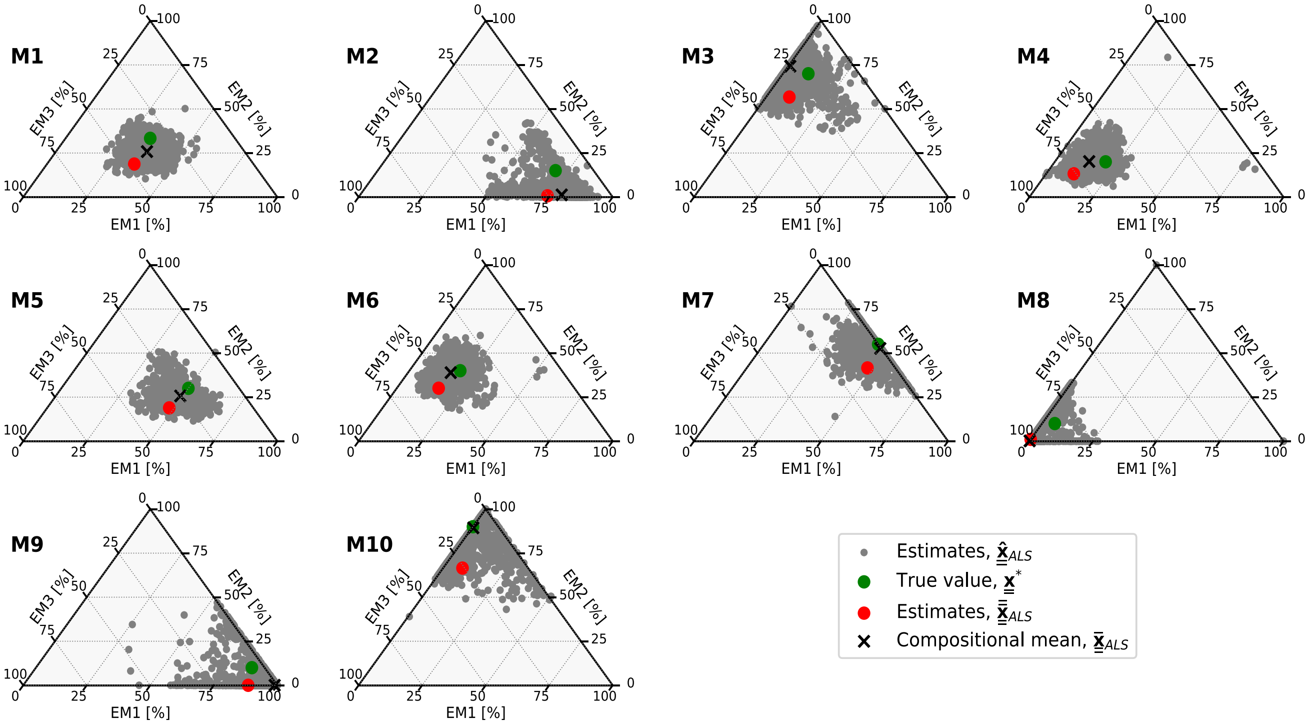

Fig. 9. Results of the production allocation approach obtained through the ALS approach. Each ternary diagram corresponds to a given mixture and includes: (i) Individual Estimates, $\hat{\underline{\mathbf{x}}}_{ALS}$), (ii) Compositional mean of $\hat{\underline{\mathbf{x}}}_{ALS}$, $\bar{\underline{\mathbf{x}}}_{ALS}$; (iii) True value of the mass fractions in the mixtures, $\underline{\mathbf{x}}^*$; and (iv) Estimates obtained by averaging mixtures and EMs replicates before performing the deconvolution, $\bar{\bar{\underline{\mathbf{x}}}}_{ALS}$.

4. Conclusions

We introduce an original deconvolution approach for production allocation to enable effective assessment of the diverse oil types forming a mixture originating from the common practice of commingling oils associated with diverse reservoirs, wells, and/or fields. Our original approach (which we term PGM) (a) is inspired by methods resting on peak ratios and (b) does not require relying on synthetic mixtures, thus being potentially associated with reduced laboratory analyses efforts. The approach is framed in the context of typically used deconvolution algorithms, i.e., the algorithm proposed by McCaffrey et al. [29], the method of Nouvelle et al. [28], as well as the approach based on the Alternating Least Square (ALS) algorithm. We also present extensions of (a) the method proposed by Nouvelle et al. [28] and (b) the ALS

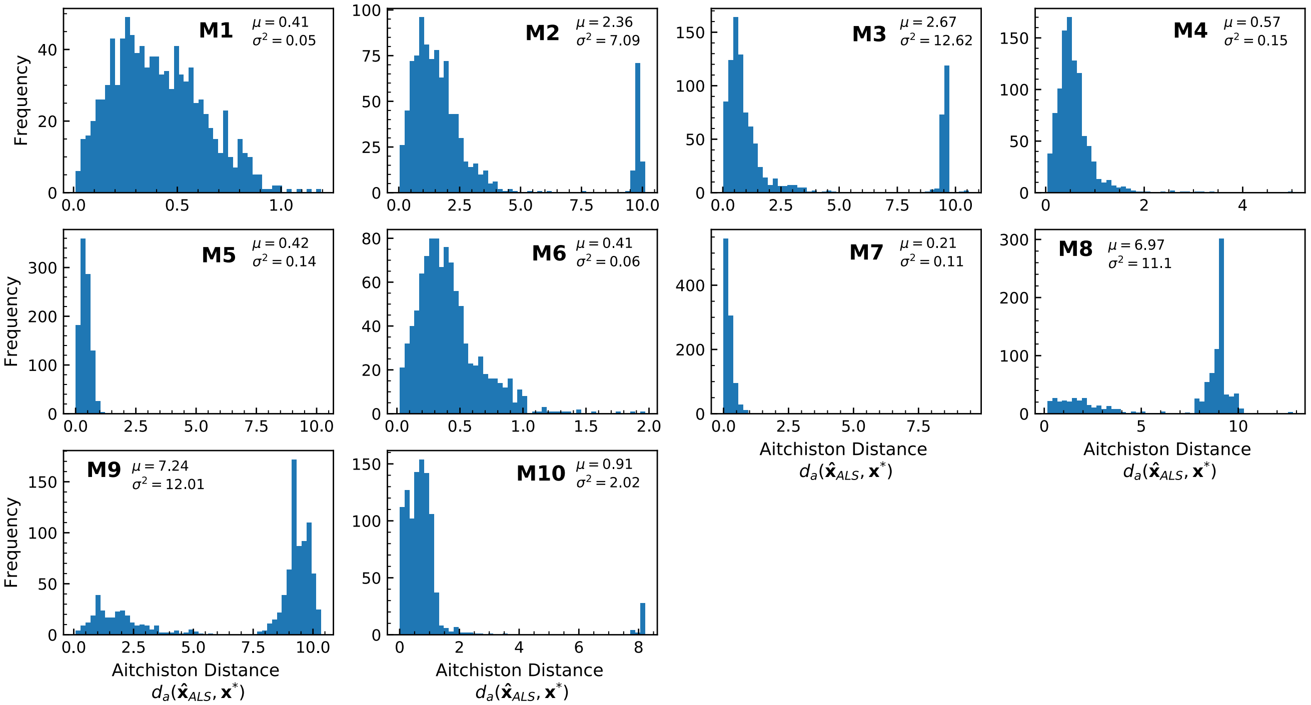

Fig. 10. Histograms of the Aitchison distances between measured mass fractions $\underline{\mathbf{x}}^*$ and rows of $\underline{\hat{\mathbf{x}}}_{ALS}$. Values of the resulting mean (μ) and variance (σ^2) are listed.

Table 3
Mean absolute error, MAE_ξ, of the deconvolution algorithms. The algorithms are implemented for two diverse numbers of features of the mixtures and of the EMs.

Error metric	Feature setting	McCaffrey	Nouvelle based on Eq. (4)	Nouvelle based on Eq. (7)	Original PGM Approach	ALS
MAE_ξ	11 Peaks (39 ratios)	8.5%	20.9%	6.7%	10%	11.7%
MAE_ξ	41 Peaks (189 ratios)	2.0%	11.6%	0.8%	0.7%	9.4%
$MAPE_\xi$	11 Peaks (39 ratios)	43.9%	109.2%	49%	48.7%	53.8%
$MAPE_\xi$	41 Peaks (189 ratios)	14.5%	75.8%	5.6%	2.3%	32.6%

algorithm, which we view in a stochastic context, corresponding to a Monte Carlo framework, with the aim of improving their robustness and reliability.

The potential of the new PGM approach is shown together with an assessment of the other analyzed deconvolution algorithms against a suite of new laboratory-based three-oil commingling scenarios. These are based on the design and introduction of a novel and low-cost experimental approach. The latter rests on a direct quantitative determination of C_8-C_{12} alkylbenzene components in oil through GC–MS fingerprinting and has been developed to circumvent some limitations of the typically employed methodologies.

Results of the analyses of the controlled experiments provide a unique, comprehensive, and rigorous comparison of the traditional production allocation deconvolution algorithms and highlight the benefit of our extensions to these and of the new PGM approach and algorithm. Our study documents that the number of features used during a quantitative deconvolution is critical to enhance the accuracy of the procedure. Additionally, we found that our new PGM approach is the most accurate methodology, followed by the Nouvelle algorithm based on our modified objective function (Eq. (7)).

CRediT authorship contribution statement

Leonardo Sandoval: Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Monica Riva:** Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Placido Franco:** Conceptualization, Methodology, Validation, Investigation, Resources, Writing – original draft. **Ivo Colombo:** Conceptualization, Methodology, Software, Resources, Writing – original draft, Project administration. **Roberto Galimberti:** Conceptualization, Methodology, Investigation, Resources, Supervision, Project administration, Funding acquisition. **Alberto Guadagnini:** Methodology, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

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.

Acknowledgment

The authors acknowledge financial and technical support from Geolog srl.

References

[1] England WA. Reservoir geochemistry – A reservoir engineering perspective. J Petrol Sci Eng 2007;58(3–4):344–54. https://doi.org/10.1016/j.petrol.2005.12.012.

[2] Cubitt JM, England WA, Larter SR. Understanding petroleum reservoirs: Towards an integrated reservoir engineering and geochemical approach. Geological Society of London 2004. https://doi.org/10.1144/GSL.SP.2004.237.

[3] van Bergen PF, Gordon M. Production geochemistry: fluids don't lie and the devil is in the detail. Geological Society, London, Special Publications 2020;484(1):9–28. https://doi.org/10.1144/SP484.1.

[4] Koolen HHF, Gomes AF, de Moura LGM, Marcano F, Cardoso FMR, Klitzke CF, et al. Integrative mass spectrometry strategy for fingerprinting and tentative structural characterization of asphaltenes. Fuel 2018;220:717–24.

[5] Yang C, Yang Z, Zhang G, Hollebone B, Landriault M, Wang Z, et al. Characterization and differentiation of chemical fingerprints of virgin and used lubricating oils for identification of contamination or adulteration sources. Fuel 2016;163:271–81. https://doi.org/10.1016/j.fuel.2015.09.070.

[6] Rudyk S, Spirov P, Sogaard E. Application of GC–MS chromatography for the analysis of the oil fractions extracted by supercritical CO2 at high pressure. Fuel 2013;106:139–46. https://doi.org/10.1016/j.fuel.2012.12.004.

[7] Permanyer A, Douifi L, Lahcini A, Lamontagne J, Kister J. FTIR and SUVF spectroscopy applied to reservoir compartmentalization: a comparative study with gas chromatography fingerprints results. Fuel 2002;81(7):861–6. https://doi.org/10.1016/S0016-2361(01)00211-3.

[8] Ekpo BO, Essien N, Neji PA, Etsenake RO. Geochemical fingerprinting of western offshore Niger Delta oils. J Petrol Sci Eng 2018;160:452–64. https://doi.org/10.1016/j.petrol.2017.10.041.

[9] Milkov AV, Goebel E, Dzou L, Fisher DA, Kutch A, McCaslin N, et al. Compartmentalization and time-lapse geochemical reservoir surveillance of the Horn Mountain oil field, deep-water Gulf of Mexico. AAPG Bull 2007;91(6): 847–76. https://doi.org/10.1306/01090706091.

[10] Chuparova E, Kratochvil T, Kleingeld J, Bilinski P, Guillory C, Bikun J, et al. Integration of time-lapse geochemistry with well logging and seismic to monitor dynamic reservoir fluid communication: Auger field case-study, deep water Gulf of Mexico. Geological Society of London, Special Publications 2010;347(1):55–70. https://doi.org/10.1144/SP347.5.

[11] Kanshio S. A review of hydrocarbon allocation methods in the upstream oil and gas industry. J Petrol Sci Eng 2020;184:1–13. https://doi.org/10.1016/j.petrol.2019.106590.

[12] McCaffrey MA, Baskin DK, Patterson BA, Ohms DH, Stone C, Reisdorf D. Oil fingerprinting dramatically reduces production allocation costs. World Oil 2012; 55.

[13] Elsinger RJ, Leenaarts EM, Kleingeld JC, van Bergen P, Gelin F. Otter-Eider Geochemical Production Allocation: 6+ Years of Continuous Monitoring to Provide Fiscal Measurements for Hydrocarbon Accounting. In: HEDBERG Conference. 2010.

[14] Patience R, Bastow M, Fowler M, Moore J, Barrie C. The Application of Petroleum Geochemical Methods to Production Allocation of Commingled Fluids. In: SPE Europec featured at 82nd EAGE Conference and Exhibition. 2021. https://doi.org/10.2118/205130-MS.

[15] Murray AP, Peters KE. Quantifying multiple source rock contributions to petroleum fluids: Bias in using compound ratios and neglecting the gas fraction. AAPG Bull 2021;105(8):1661–78. https://doi.org/10.1306/03122120056.

[16] Carati C, Bonoldi L, Bonetti R, Nali M, Amendola A. Production Allocation of Commingled Reservoir Fluids by On-Site Spectroscopic Analysis. In: International Petroleum Technology Conference. 2020. https://doi.org/10.2523/IPTC-20057-MS.

[17] Yang W, Casey JF, Gao Y, Li J. A new method of geochemical allocation and monitoring of commingled crude oil production using trace and ultra-trace multi-element analyses. Fuel 2019;241:347–59. https://doi.org/10.1016/J.FUEL.2018.12.049.

[18] Peters KE, Ramos LS, Zumberge JE, Valin ZC, Bird KJ. De-convoluting mixed crude oil in Prudhoe Bay field, North Slope. Alaska. Organic Geochemistry 2008;39(6): 623–45. https://doi.org/10.1016/j.orggeochem.2008.03.001.

[19] Hwang RJ, Baskin DK, Teerman SC. Allocation of commingled pipeline oils to field production. Org Geochem 2000;31(12):1463–74. https://doi.org/10.1016/S0146-6380(00)00123-6.

[20] Kaufman RL, Ahmed AS, Hempkins WB. A new technique for the analysis of commingled oils and its application to production allocation calculations. In: 16th Annual Convention of the Indonesian Petroleum Asociation. 1987.

[21] Kaufman RL. Gas chromatography as a development and production tool for fingerprinting oils from individual reservoirs: applications in the Gulf of Mexico. In: GCSSEPM Foundation Ninth Annual Research Conference Proceedings. 1990. https://doi.org/10.5724/gcs.90.09.0263.

[22] Mohamed MS. Obaiyed field fluid Geochemical analysis. In: Abu Dhabi International Petroleum Exhibition and Conference. 2000. https://doi.org/10.2118/87289-MS.

[23] Jweda J, Michael E, Jokanola OA, Hofer R, Parisi VA. Optimizing field development strategy using time-lapse geochemistry and production allocation in Eagle Ford. In: SPE/AAPG/SEG Unconventional Resources Technology Conference. 2017 https://doi.org/10.15530/URTEC-2017-2671245.

[24] Liu F, Michael E, Johansen K, Brown D, Allwardt J. Time-lapse geochemistry (TLG) application in unconventional reservoir development. In: Unconventional Resources Technology Conference. 2017 https://doi.org/10.15530/urtec-2017-2670186.

[25] Barrie CD, Donohue CM, Zumberge JA, Zumberge JE. Production Allocation: Rosetta Stone or Red Herring? Best Practices for Understanding Produced Oils in Resource Plays. Minerals 2020;10(12):1105. https://doi.org/10.3390/min10121105.

[26] Zhan Z-W, Tian Y, Zou Y-R, Liao Z, Peng P. De-convoluting crude oil mixtures from Palaeozoic reservoirs in the Tabei Uplift, Tarim Basin. China. Organic Geochemistry 2016;97:78–94.

[27] Baskin DK, Kornacki A, McCaffrey M. Allocating the Contribution of Oil from the Eagle Ford Formation, the Buda Formation, and the Austin Chalk to Commingled Production from Horizontal Wells in South Texas Using Geochemical Fingerprinting Technology. Search & Discovery 2013;41268.

[28] Nouvelle X, Rojas K, Stankiewicz A. Novel method of production back-allocation using geochemical fingerprinting. In: Abu Dhabi International Petroleum Conference and Exhibition. 2012. https://doi.org/10.2118/160812-MS.

[29] McCaffrey MA, Ohms DS, Werner M, Stone CL, Baskin DK, Patterson BA. Geochemical allocation of commingled oil production or commingled gas production. In: SPE Western North American Region Meeting. 2011. https://doi.org/10.2118/144618-MS.

[30] Spagnolini U. Statistical Signal Processing in Engineering. John Wiley & Sons; 2018.

[31] Ljung L. Perspectives on system identification. Annual Reviews in Control 2010;34 (1):1–12.

[32] Li L, Tan J, Wood DA, Zhao Z, Becker D, Lyu Q, et al. A review of the current status of induced seismicity monitoring for hydraulic fracturing in unconventional tight oil and gas reservoirs. Fuel 2019;242:195–210. https://doi.org/10.1016/j.fuel.2019.01.026.

[33] Dembicki H. Practical petroleum geochemistry for exploration and production. Elsevier; 2016.

[34] Amendola A, Caldiero L, Cerioli Regondi AMA, Dolci D, Galimberti R, Nali M. Production Allocation Without End Members: Now It Is Possible. In: Offshore Mediterranean Conference and Exhibition. 2017.

[35] Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett 1999;9(3):293–300. https://doi.org/10.1023/A:1018628609742.

[36] Johansson D, Lindgren P, Berglund A. A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. Bioinformatics 2003;19(4):467–73. https://doi.org/10.1093/bioinformatics/btg017.

[37] Yeniay Ö, Göktaş A. A comparison of partial least squares regression with other prediction methods. Hacettepe Journal of Mathematics and Statistics 2002;31: 99–111.

[38] Wang H, Chu X, Chen P, Li J, Liu D, Xu Y. Partial least squares regression residual extreme learning machine (PLSRR-ELM) calibration algorithm applied in fast determination of gasoline octane number with near-infrared spectroscopy. Fuel 2022;309:122224. https://doi.org/10.1016/j.fuel.2021.122224.

[39] de Lima FW, Corgozinho CN, Tauler R, Sena MM. Monitoring biodiesel and its intermediates in transesterification reactions with multivariate curve resolution alternating least squares calibration models. Fuel 2021;283:119275. https://doi.org/10.1016/j.fuel.2020.119275.

[40] Bao X, Dai L. Partial least squares with outlier detection in spectral analysis: A tool to predict gasoline properties. Fuel 2009;88(7):1216–22. https://doi.org/10.1016/j.fuel.2008.11.025.

[41] Bianchi Janetti E, Dror I, Guadagnini A, Riva M, Berkowitz B. Estimation of single-metal and competitive sorption isotherms through maximum likelihood and model quality criteria. Soil Sci Soc Am J 2012;76(4):1229–45.

[42] Jaumot J, de Juan A, Tauler R. MCR-ALS GUI 2.0: New features and applications. Chemometrics and Intelligent Laboratory Systems 2015;140:1–12. https://doi.org/10.1016/j.chemolab.2014.10.003.

[43] Brunton SL, Data-Driven KJN. Science & Engineering. Machine Learning, Dynamical Systems, and Control. Cambridge University Press; 2019.

[44] Llinares E, Igual J, Camacho A. Application of regularized Alternating Least Squares to an astrophysical problem. Appl Math Comput 2012;219:1367–74. https://doi.org/10.1016/j.amc.2012.07.044.

[45] Menafoglio A, Guadagnini L, Guadagnini A, Secchi P. Object oriented spatial analysis of natural concentration levels of chemical species in regional-scale aquifers. Spatial Statistics 2021;43:100494. https://doi.org/10.1016/j.spasta.2021.100494.

[46] Menafoglio A, Guadagnini A, Secchi P. A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. Stoch Env Res Risk Assess 2014;28:1835–51. https://doi.org/10.1007/s00477-014-0849-8.

[47] van den Boogaart KG, Tolosana-Delgado R. Analyzing Compositional Data with R, Analyzing Compositional Data with R. Springer.