Calibration model averaging in chemical analysis: a case study for the method of standard additions

Pagliano, Enea; Meija, Juris

Publisher's version / Version de l'éditeur:
https://doi.org/10.1088/1681-7575/accd75
Metrologia, 60, 3, 2023-05-11

OPEN ACCESS
IOP Publishing | Bureau International des Poids et Mesures Metrologia
Metrologia 60 (2023) 035003 (6pp) https://doi.org/10.1088/1681-7575/accd75

Calibration model averaging in chemical analysis: a case study for the method of standard additions

Enea Pagliano and Juris Meija*

National Research Council Canada, 1200 Montreal Road, Ottawa ON K1A0R6, Canada

E-mail: juris.meija@nrc-cnrc.gc.ca

Received 26 October 2022, revised 3 April 2023
Accepted for publication 17 April 2023
Published 11 May 2023

Abstract

In this study we demonstrate that the method of standard additions can provide unreliable results in chemical analysis when the linear calibration model is forced to fit the data. A well designed standard addition experiment can still yield results biased by 10% when the analyst relies only on the linear model. Recently, the Joint Committee for Guides in Metrology (GUM-6) has emphasized the need to address the uncertainty inherent to the choice of measurement models and here we show how model averaging can provide a practical way to account for model uncertainty in the method of standard additions.

Supplementary material for this article is available online

Keywords: method of standard additions, uncertainty evaluation, model averaging

(Some figures may appear in colour only in the online journal)

1. Introduction

Measurements are model-based inferences [1], yet the importance of the measurement model equation is often overlooked. Indeed, while one may not be surprised that various measurement models can lead to different results, often only a single 'best' model is utilized which carries the impression that it is the only model that can explain the data. Given that calibration models in chemical analysis are chosen empirically, such an approach fails to incorporate the uncertainty due to model choice and, in turn, can lead to over-confident predictions [2]. The method of standard additions has been a staple in analytical chemistry for nearly a century and is often thought to provide more accurate measurement results than the application of external calibration [3]. Out of simplicity, the linear calibration model ($y = a + bx$) is often adopted [4, 5]. However, such an approach has several limitations which we discuss in this study.

Consider a motivational example to determine nitrate ions in a standard solution with a known mass fraction $w(NO_3^-) = (50.5 \pm 0.2)\ \mathrm{mg\,kg^{-1}}$. Five different amounts of nitrate standard solution (having natural isotopic composition) were added to the sample aliquots along with a fixed amount of isotopically labelled internal standard ($^{15}NO_3^-$). The mass fraction of nitrate ions in the resulting mixtures was measured by electron impact ionization GC-MS after derivatization (ethylation) with triethyloxonium ions [6, 7]. The standard-addition calibration curve (figure 1) was obtained by plotting the mass fraction of nitrate standard added to the sample aliquots (x-axis) against the normalized analytical signal (y-axis) which is the chromatographic peak area of $EtO^{14}NO_2$ derivative divided by the peak area of $EtO^{15}NO_2$. The experimental design adopted in this example follows the conventional approach where the

* Author to whom any correspondence should be addressed.

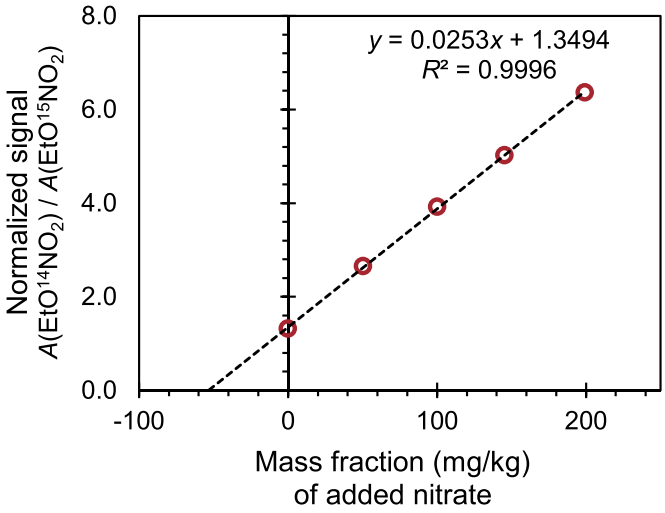

Figure 1. Conventional method of standard additions to determine nitrates in a standard solution. The linear calibration model with ordinary least squares fitting provides a result $w(\mathrm{NO_3^-}) = (53.3 \pm 1.8)$ mg kg^{-1} that is inconsistent with the known value $w(\mathrm{NO_3^-}) = (50.5 \pm 0.2)$ mg kg^{-1} (95% confidence intervals quoted).

Table 1. Motivational example of nitrate determination by GC-MS using the method of standard additions[a].

Level	Normalized signal	Mass fraction (mg kg^{-1}) of added nitrate
0	1.313, 1.308, 1.319	0
1	2.645, 2.659, 2.647	50.46(10)
2	3.872, 3.969, 3.911	100.20(20)
3	5.003, 5.013, 5.045	145.27(30)
4	6.352, 6.342, 6.394	199.19(40)

[a] The normalized signal is the peak area corresponding to the nitrate derivative ($\mathrm{EtO^{14}NO_2}$) divided by the response of the isotopically labeled internal standard ($\mathrm{EtO^{15}NO_2}$). Standard uncertainty associated with the mass fraction of added nitrate (applicable to the last two digits) is given in parentheses.

analytical signal of the highest spiked sample is at least four times that of the unspiked sample [4].

As shown in figure 1, the measurement results exhibit high precision and appear to follow a linear relationship. Using the measurement data from table 1, the mass fraction of nitrate in the standard solution is obtained as the ratio of the intercept and the slope of the linear regression between the analytical signal and the mass fraction of added nitrate. The ordinary least squares (OLS) fit provides a result $w = 53.3$ mg kg^{-1} with the associated standard uncertainty $u(w) = 0.9$ mg kg^{-1}. The choice of OLS is common because the uncertainty associated with the mass fraction of added standard is negligible. Indeed, a more rigorous Bayesian errors-in-variables linear regression model gives the same estimate along with a similar uncertainty, $u(w) = 1.0$ mg kg^{-1} (paragraph S1). Both of these results are significantly different from the certified value of nitrate in this sample and thus illustrate the dangers in relying on a single empirical measurement model.

The limitations of forcing a linear response model to standard addition curves have been noted before. However, the proposed solutions to address these shortcomings have been to employ either polynomial [8] or rational [9] calibration functions. Here we build on this prior work and propose using *all* of these measurement models to address the limitations that any single empirical calibration model can impose in practice.

In addition to the choice of measurement model, experimental design also has a significant effect on the results and in this work we adopted experimental designs which are commonly used in practice [4].

2. Model averaging

Measurement models are often empirical, meaning that they are chosen for their performance. The use of polynomial functions is an example of such a measurement model [8]. The problem often faced in analytical method development is to decide what model to choose among the several viable candidates. Commonly, one selects the measurement model based on the goodness-of-fit with some penalty added for the number of parameters in the model (k) and the amount of data points at hand (n). Such a measure, B, is known as the Bayesian information criterion (BIC) [10]. Assuming Gaussian errors, B can be derived from the familiar sum squares of regression residuals,

$$s_i^2 = \frac{1}{n} \sum_{j=1\ldots n} \left(y_j - f_i(x_j)\right)^2 \tag{1}$$

and takes the following form [11]:

$$B(f_i) = -2L_i + (1 + k_i)\ln(n) \tag{2}$$

where L_i is the maximum value for the model log-likelihood [11]:

$$L_i = -\frac{n}{2}\ln\left(2\pi\,\mathrm{e}\,s_i^2\right). \tag{3}$$

The calculation of B is greatly facilitated in R using function `BIC` defined in package `stats` [12].

A model with the lowest value of B is typically selected as the measurement model. However, instead of selecting a single best model to obtain the measurement result (w), one can take the average result from *all* candidate models, each with a probability (p_i) that is proportional to how well it fits the data (x, y):

$$w = \sum_i p_i f_i(x, y). \tag{4}$$

If all candidate models f_i are considered with equal prior probability, then the statistical model weights are calculated from B as follows [13]:

$$p_i = \frac{\exp\left(-\frac{1}{2}B(f_i)\right)}{\sum_j \exp\left(-\frac{1}{2}B(f_j)\right)}. \tag{5}$$

Model averaging is useful when there is uncertainty about the exact nature of the data generation process. Indeed, this general approach is adopted widely in science [14] with a recent example involving radiocarbon dating of cultural artefacts from Machu Picchu [10, 15]. There, the calibration models for radiocarbon dating typically applied to samples from northern (IntCal) or southern (SHCal) hemispheres were averaged as a way to recognize the fact that large parts of the Eastern Andes have atmospheric carbon influx from both hemispheres.

3. Results and discussion

This section provides several examples aimed at illustrating the model averaging in evaluating the results of standard additions. All calculations were performed using R, the free software environment for statistical computing (v.4.0.2) [12], and the source code is provided in the supporting information.

In this proof-of-concept study, four calibration models were used for the interpretation of standard additions data:

$$\begin{aligned} \text{linear} \quad & x = a + by, \\ \text{rational} \quad & x = (a + by)/(1 + cy), \\ \text{quadratic} \quad & x = a + by + cy^2, \\ \text{reduced cubic} \quad & x = a + by + cy^3. \end{aligned}$$

The fitting was performed using ordinary linear least squares (OLS) and the coordinate swapping method was employed to obtain the results (w) [16]. Thus, y was taken as the analytical signal and x as the amount of analyte added to the sample aliquots. Under such an approach, the result is given as $w = a$ and $u(w) = u(a)$ [16]. The rational measurement function was also fitted to the data using OLS method by using the linearized model expression $x = a + by - cxy$ [17].

When the uncertainty of the analytical signals was available, the Monte Carlo method [18] was used to evaluate the uncertainty of the standard additions results. In particular, the model averaging calculation was performed 10 000 times, each time re-sampling the analytical signal data (y) from a Gaussian distribution with mean y and standard deviation $u(y)$. Furthermore, results from each measurement model were further re-sampled to include the uncertainties of the obtained model parameters.

3.1. In silico *data: linear response*

Consider the following five-level standard addition data set: x = (0, 1, 2, 3, 4) mg kg^{-1}, y = (1, 2, 3, 4, 5), and $u(y)$ = 0.01 for all measured signals (homoscedastic response). The underlying response model is linear with the true value of analyte $w = 1$ mg kg^{-1}. As expected, the linear calibration model yields unbiased result, w = (1.000 ± 0.030) mg kg^{-1} (figure S1, table S1) with similar results from the model averaging: w = (1.000 ± 0.050) mg kg^{-1}. We also obtain similar results with the analysis of bromide ions in BCR-611 reference material by GC-MS (figure S2, table S2) [19].

It is evident from this example that model averaging incurs an additional uncertainty and that is the price one must pay if there are doubts about the true response model.

3.2. In silico *data: non-linear response*

Linear response (i.e. the instrumental signal is directly proportional to the mass fraction of analyte) remains an idealized concept in modern analytical chemistry methods. Indeed, even the best detectors will display deviations from linear response at very low or very high concentrations which is encountered in UV–vis spectrophotometry. For example, Li *et al* reports deviations from the classical Beer–Lambert law between the 216 nm and 230 nm light absorbance (y) and the concentration of SO_2 (x) using the following empirical equation (we omit the measurement units) [20]:

$$y = 0.111 \cdot x + 275 \cdot \log\left(\frac{2}{1 + \exp(0.000727 \cdot x)}\right). \quad (6)$$

We shall consider an *in silico* standard additions experiment where x = (0, 150, 300, 450, 600) mg l^{-1} and the corresponding analytical signals are obtained from equation (6) whose rounded values were y = (9.96, 19.57, 28.82, 37.73, 46.31). A small uncertainty, $u(y) = 0.10$, is assigned to all analytical signals and the true value is set to w = 150 mg l^{-1}.

As shown in the SI (figure S3, table S3), despite the seemingly good fit, the linear calibration model fails spectacularly: the result of the linear calibration model (w = (170 ± 14) mg l^{-1}) is 3-σ (10%) away from the true value. In contrast, the weighted average of the results from all measurement models considered here, w = (151 ± 10) mg l^{-1}, is a faithful estimate of the true value.

This example demonstrates that the application of the linear calibration model to standard addition data that are intrinsically non-linear can result in large biases.

3.3. *Motivational example*

The motivational example described in the introduction is the experimental counterpart of the *in silico* example with non-linear response. This is so because the use of isotopic internal standards leads to a non-linear response [17, 21]. Thus, the calibration curve shown in figure 1 is intrinsically non-linear even though its non-linear nature is hardly noticeable.

Figure 2 shows that the selected non-linear models perform better than the linear model. Furthermore, the weighted average of all models (w = (49.5 ± 3.3) mg kg^{-1}) agrees well with the known gravimetric value, (50.5 ± 0.2) mg kg^{-1} NO_3^-.

In this case, the true response model is the rational function, $x = (a + by)/(1 + cy)$ [17, 21], so it is not surprising that the model averaging has assigned the largest weight (50%) to this

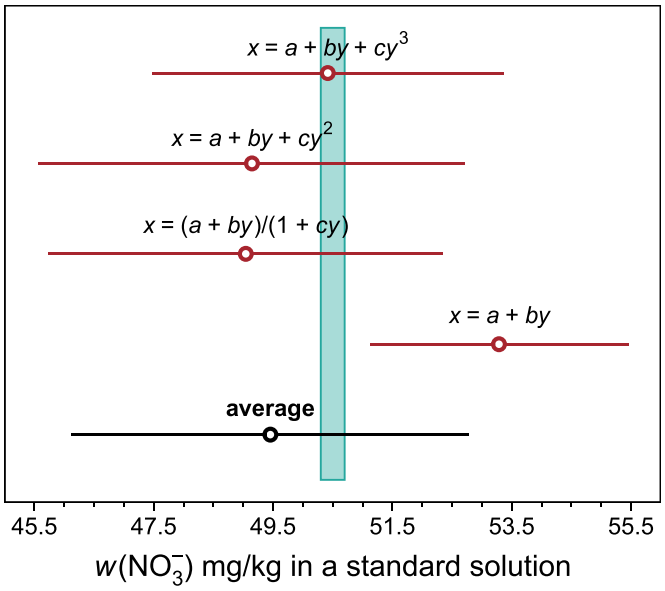

Figure 2. Performance of various calibration models in the measurement of nitrates by GC-MS using the method of standard addition (figure 1 and table 1). The error bars represent 95% coverage intervals and the blue rectangle represents the certified value and its uncertainty.

model whereas giving only 5% to the linear calibration model (table S4).

3.4. Nitrites in seawater

The previous examples describe a situation where the linear model provides a result that is approx. 5% from the true value which might be acceptable to many analysts.

Consider standard addition experiment where nitrite ions are determined in seawater by spectrophotometry using the Griess method [22, 23]. A NO_2^- standard solution was prepared gravimetrically in a low nutrient seawater (OSIL, UK) to yield $w(NO_2^-) = (0.3657 \pm 0.0015)$ mg kg^{-1}. This standard solution was then treated as a sample and its NO_2^- levels were determined using five-level standard addition method (figure 3). All absorbance readings were in the interval [0.4, 0.9] where the deviations from the Beer–Lambert law are usually negligible.

The OLS fitting with the linear calibration model shows no clear patterns in the regression residuals suggesting suitability of a linear model (figure S12). Despite the seemingly linear response, the standard addition result is biased by more than 10% (see table S5). The non-linear calibration models, however, provide unbiased estimates and so does the model average.

This example shows that even a seemingly good experimental design leading to a linear regression with no patterns detected from the residual analysis is not sufficient to ensure accurate results from standard additions. Indeed, the results

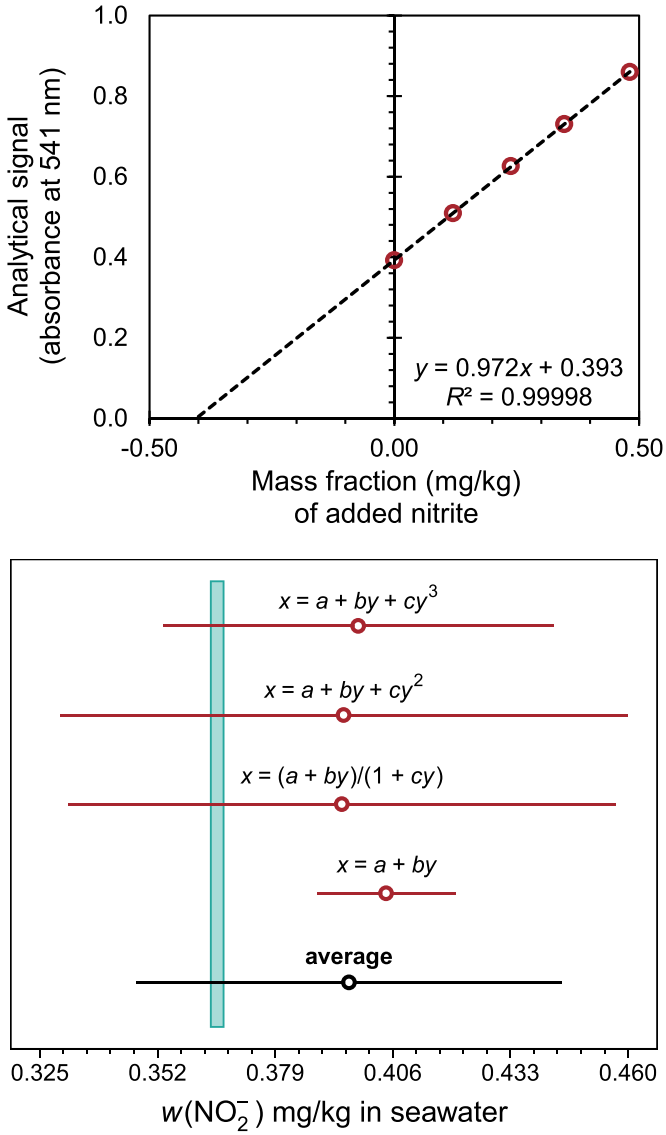

Figure 3. Determination of nitrite ions in seawater by spectrophotometry using the method of standard additions. Top: the standard addition calibration plot with the linear model. Bottom: comparing the performance of different measurement models. Error bars represent 95% coverage intervals and the blue rectangle represents true value and its uncertainty.

obtained from the traditional implementation of standard additions should always be scrutinized.

3.5. Nitrates in spinach

The previous example raises awareness about the magnitude of errors that can arise when the linear calibration model is the only model considered in standard addition experiments.

The consideration of model choice becomes particularly critical in high-precision measurements. Consider the certification of nitrate ions in the NRC spinach powder reference material SPIN-1 [24]. The NO_3^- content in this CRM

Table 2. Comparison of the linear calibration model and the model average for determination of nitrate ions in spinach CRM (NRC SPIN-1) by ion chromatography with conductivity detection. The certified value is (22.53 ± 0.43) mg g^{-1} NO_3^-.

	Linear model		Model average	
Sample	w, mg g^{-1}	$u(w)$	w, mg g^{-1}	$u(w)$
1	21.4	1.0	23.3	1.4
2	21.2	1.2	23.0	2.0
3	21.4	0.8	22.0	1.3
4	22.1	0.9	22.5	1.4
5	21.8	0.8	23.2	0.7
Mean bias	−4.2%		+1.2%	

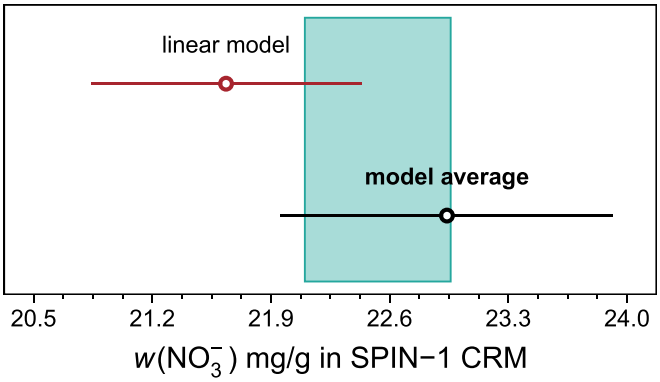

Figure 4. Determination of nitrate ions in spinach CRM by ion chromatography with conductivity detection using the method of standard additions: comparison of the linear calibration model and model average. Error bars represent 95% coverage intervals and the blue rectangle represents the certified value and its uncertainty.

was assigned by high-precision isotope dilution to $w(NO_3^-) = (22.53 \pm 0.43)$ mg g^{-1} [25]. During the certification campaign, ion-chromatography coupled with conductivity detection was used as an additional measurement method. A five-level standard additions calibration was repeated five times and the results obtained by the linear calibration model and by model average are shown in table 2 (and in figures S6–S10 and tables S6–S10.)

The results of these five independent replicates (figure 4) show, on average, a −4.2% bias when the linear calibration model is used. In contrast, model averaging yields, on average, a +1.2% bias from the certified value. While both results agree with the certified value at the 95% confidence level, model average performs markedly better.

3.6. Future improvements

In the examples presented in this work, the model average has generally provided more reliable results than the linear calibration model alone. While the model averaging resulted in lower biases throughout several examples, it is not a guarantee for unbiased performance in general. In this vein, we provide an example of phosphate measurements by spectrophotometry (figure S11, table S11) where the linear calibration model and model average both return biased results.

The performance of the model averaging can be further improved by choosing additional mathematical models that adequately explain the analytical response. Hence, the availability of theoretical or empirical equations that can describe deviations from linear response for a variety of analytical methods might prove useful for future development of this approach [20, 26]. In addition, we have employed ordinary least squares fitting and the use of a more rigorous errors-in-variables approach could be investigated.

4. Conclusions

The results of the standard additions method can be strongly influenced by the choice of the calibration model. Indeed, we have shown that the traditional implementation of the linear calibration model can yield biased results even when the calibration plots appear linear.

Through several practical examples, we show that the results obtained from model averaging are more robust than the traditional application of the linear calibration model alone. Thus, considering that the application of a single empirical calibration model can lead to unsatisfactory results, we argue that model averaging should become part of the data analysis toolkit when evaluating the results of chemical measurements.

ORCID iDs

Enea Pagliano https://orcid.org/0000-0001-9755-8737
Juris Meija https://orcid.org/0000-0002-3349-5535

References

[1] Tal E 2020 Measurement in science *The Stanford Encyclopedia of Philosophy* Fall 2020 edn, ed E N Zalta (Stanford, CA: Metaphysics Research Lab, Stanford University)
[2] Bailey D C 2017 *R. Soc. Open Sci.* **4** 160600
[3] Burns D T and Walker M J 2019 *Anal. Bioanal. Chem.* **411** 2749–53
[4] Ellison S L R and Thompson M 2008 *Analyst* **133** 992–7
[5] Hauswaldt A L, Rienitz O, Jährling R, Fischer N, Schiel D, Labarraque G and Magnusson B 2012 *Accredit. Qual. Assur.* **17** 129–38
[6] Pagliano E, Onor M, Pitzalis E, Mester Z, Sturgeon R E and D'Ulivo A 2011 *Talanta* **85** 2511–6
[7] Pagliano E, Campanella B, D'Ulivo A and Mester Z 2018 *Anal. Chim. Acta* **1025** 12–40
[8] Wieczorek M, Debosz M, Świt P, Piech A, Kasperek J and Kościelniak P 2018 *Monatsh. Chem.* **149** 1567–72
[9] Gorazda K, Michałowska-Kaczmarczyk A M, Asuero A G and Michałowski T 2013 *Talanta* **116** 927–30
[10] Possolo A and Meija J 2022 *Measurement Uncertainty: A Reintroduction* 2nd edn (Montevideo: Sistema Interamericano de Metrologia)
[11] Findley D F 1991 *Ann. Inst. Stat. Math.* **43** 505–14

[12] R Core Team 2020 R: a language and environment for statistical computing (Vienna: R Foundation for Statistical Computing) (available at: www.R-project.org/)
[13] Wit E, van den Heuvel E and Romeijn J W 2012 *Stat. Neerl.* **66** 217–36
[14] Hinne M, Gronau Q F, van den Bergh D and Wagenmakers E J 2020 *Adv. Methods Pract. Psychol. Sci.* **3** 200–15
[15] Burger R L, Salazar L C, Nesbitt J, Washburn E and Fehren–Schmitz L 2021 *Antiquity* **95** 1265–79
[16] Meija J, Pagliano E and Mester Z 2014 *Anal. Chem.* **86** 8563–7
[17] Pagliano E, Mester Z and Meija J 2015 *Anal. Chim. Acta* **896** 63–67
[18] JCGM 101 2008 *Evaluation of Measurement Data—Supplement 1 to the 'Guide to the Expression of Uncertainty in Measurement'—Propagation of Distributions using a Monte Carlo Method* (available at: https://www.bipm.org/en/committees/jc/jcgm/publications)
[19] Pagliano E, Mester Z and Meija J 2013 *Anal. Bioanal. Chem.* **405** 2879–87
[20] Li L, Zhao H, Ni N, Wang Y, Gao J, Gao Q, Zhang Y and Zhang Y 2022 *Spectrochim. Acta* A **275** 121192
[21] Pagliano E and Meija J 2021 *Int. J. Mass Spectrom.* **464** 116557
[22] Grasshoff K, Kremling K and Ehrhardt M 1999 *Methods of Seawater Analysis* 3rd edn (Weinheim: Wiley)
[23] Pagliano E, Nadeau K, Mihai O, Pihillagawa Gedara I and Mester Z 2022 *Anal. Bioanal. Chem.* **414** 4745–56
[24] Pagliano E *et al* SPIN-1: air-dried spinach powder certified reference material for Nitrate and trace metals (Ottawa: National Research Council Canada) (https://doi.org/10.4224/crm.2019.spin-1)
[25] Pagliano E *et al* 2019 *Anal. Bioanal. Chem.* **411** 3435–45
[26] Tolbin A Y, Pushkarev V E and Tomilova L G 2018 *Chem. Phys. Lett.* **706** 520–4