Accred Qual Assur (2018) 23:57–71
https://doi.org/10.1007/s00769-017-1296-2

GENERAL PAPER

Measurements recovery evaluation from the analysis of independent reference materials: analysis of different samples with native quantity spiked at different levels

Rui M. S. Cordeiro[1,2] · Constantino M. G. Rosa[2] · Ricardo J. N. Bettencourt da Silva[1]

Received: 17 January 2017 / Accepted: 11 October 2017 / Published online: 18 November 2017

Abstract Measurement uncertainty evaluation involves combining uncertainty components reflecting all relevant random and systematic effects: the precision and trueness uncertainty components, respectively. Typically, trueness is assessed through the analysis of various materials with known reference value, such as certified reference materials (CRMs) or spiked samples, from which it should be decided about the relevance and the need to correct measurement results for systematic effects. Algorithms proposed so far to assess systematic effects are only applicable to the analysis of the same reference material type or assume that some uncertainty components affecting evaluations are negligible or constant. This work presents detailed algorithms for the assessment of systematic effects, through the determination of recovery and the respective recovery uncertainty, applicable to the analysis of various independent reference materials, such as CRMs and spiked samples with native analyte. These algorithms are applicable to cases where native analyte and/or spiking values are associated with relevant and significantly different uncertainties allowing for a reliable assessment of systematic effects and measurement uncertainty for these complex cases. This methodology was successfully applied to the quantification of Na, K, Mg, Ca, Cr, Mn, Fe and Cu in water samples from two proficiency testing schemes, by ICP-OES, where recovery was estimated from the analysis of samples with different native concentrations and spiked at different levels. The relative expanded uncertainties of the measurement results ranged from 28.9 % to 3.9 % and are fit for the monitoring of environmental water samples in accordance with criteria set in the European Union legislation.

Keywords Recovery · Uncertainty · Validation · ICP-OES · Metals · Water

✉ Ricardo J. N. Bettencourt da Silva
rjsilva@fc.ul.pt

[1] CQE – Centro de Química Estrutural, Faculdade de Ciências da Universidade de Lisboa Edifício C8, Campo Grande, 1749-016 Lisbon, Portugal

[2] Labelec – EDP, Rua Cidade de Goa 4, 2685-039 Sacavém, Portugal

Introduction

The evaluation of measurement uncertainty aims at estimating the impact of all analytical steps and effects that contribute to the measurement error (i.e. the difference between the measured and the reference quantity values [1]) in order to produce an interval that should encompass the conventional true value of the measurand with a known probability. The effects contributing to the measurement uncertainty can be divided into random and systematic effects.

The generic term ‘quantity’ is used when concepts are applicable to various specific quantities such as mass, concentration, mass concentration, mass fraction, pH and conductivity.

Different approaches have been developed to estimate the measurement uncertainty that use different types of information specific to the implementation of the measurement procedure in the laboratory or applicable to several laboratories [2–4]. For most analytical applications, the selected approach for the evaluation of the measurement uncertainty is the simplest one to apply that guarantees that the reported uncertainty is smaller than the target (i.e. maximum admissible) uncertainty [1, 5, 6].

The more pragmatic approach for the evaluation of the measurement uncertainty based on the specific performance of a laboratory, collected during the in-house validation of the measurement procedure, divides uncertainty components into precision, trueness and other components. Some authors designate the trueness component as the bias component. This approach is designated 'supra-analytical' [3], 'single-laboratory validation' [4] or 'top-down based on in-house validation data'. The trueness uncertainty component is also relevant for more detailed models of the measurement uncertainty such as the ones produced by the differential approach [7–9].

The precision and trueness uncertainty components are usually dominated by random and systematic effects, respectively. For instance, the standard deviation of the intermediate precision used to quantify the precision component reflects the randomisation of systematic effects attributed to the daily run of analysis. The trueness component is also not a 'pure' representation of systematic effects since it is not possible to perform an infinite number of replicate measurements that would produce a mean not affected by random effects.

The trueness measurement uncertainty can be estimated from results of the analysis of internal and/or external reference materials. External reference materials, such as certified reference materials or proficiency test materials, are ideal references if the analyte speciation or bonding in the matrix of the reference materials is analytically similar to the analyte speciation or bonding in the matrix of the samples to be analysed. The reference value should be traceable to an adequate reference, typically an SI unit, and have an uncertainty smaller than one-fifth of the target uncertainty to make it easier to produce measurements with an uncertainty smaller than the target value.

If no external reference material is available, the analysis of spiked materials can allow the assessment of the systematic effects. The spiking can be performed on items with or without native quantity. A native quantity is a quantity present in the analysed item, i.e. not artificially added/spiked to the analysed item in the laboratory. The native analyte is typically present from the natural cycle of analyte occurrence (e.g. the contamination process of heavy metals in river water). The spiking of materials without detectable levels of the native quantity in the material allows for the assessment of measurement performance with a smaller uncertainty since the additional uncertainty component associated with the quantification of the native quantity is eliminated. However, in some fields it is not possible to have 'real' materials free from the quantity of interest, such as oranges without ascorbic acid or urban wastewaters without nitrates. The analysis of spiked samples has the advantage of testing performance in laboratory samples, and these materials are cheaper than external references. However, if the analyte speciation and/or bonding to the matrix is critical for measurement performance, the spiking reference and methodology must be carefully selected. In many cases, the spiking reference is a stock solution of the analyte from which a portion is taken to be added to an aliquot of the studied matrix. The spiking methodology describes how the reference is added to the matrix including procedures that try to promote the interaction of the spike with the matrix such as a delay of some hours between spiking and analysis to allow some interaction between added quantity and the matrix. For instance, in the analysis of total mercury in fish tissue, sample preparation can volatilise the naturally occurring methylmercury more easily than spiked inorganic mercury. Therefore, fish tissue should be spiked with methylmercury instead of mercury(II) nitrate reference solution.

Since the magnitude of systematic effects is frequently proportional to the quantity of interest, their value is monitored by the value of the ratio of the estimated and the reference quantity value of the reference material known as 'recovery'. The reference value should be adequate for the studied measurement; e.g. if the aqua regia extractable mass fraction of chromium in a soil is measured, the reference mass fraction of total chromium in soil is inadequate if only a fraction of total chromium is extracted by the aqua regia.

The statistical and metrological quality of the estimated recovery increases if the mean of various recovery values (i.e. the mean recovery) is estimated from the analysis of the same or difference reference materials. A mean recovery close to 1 or 100 % suggests that the estimated values are not affected by recovery. The mean recovery is less affected by random effects as the number of estimated recoveries increases [10]. If at least 25 recovery tests are performed, the random effects affecting mean recovery estimation are at least five times less than the ones affecting single measurements, making it negligible in the measurement uncertainty evaluation for a single measurement result. (The standard deviation, $s(\bar{x})$, of a mean of 25 results is $\sqrt{25} = 5$ times smaller than the standard deviation, s, of a single measurement: $s(\bar{x}) = s/\sqrt{25}$ [10].) Regardless of mean recovery uncertainty relevance, the mean recovery should be used to correct results affected by large or low mean recovery values, if necessary.

The systematic effects can also be quantified by the mean relative error (i.e. the mean of ratios between measurement error and the reference value). The mean relative error can be estimated by subtracting the mean recovery by one.

After the mean recovery has been estimated, it is necessary to assess whether any deviation to the ideal 100 % recovery is relevant.

Some authors proposed combining the mean error [11] or the mean squared error [12] with the expanded or squared standard uncertainty of results not corrected for recovery, respectively, to avoid the need to assess recovery magnitude. These combinations are suggested for operationally defined measurands/measurement procedures or for cases where estimated mean error has a chance of not being representative of performance in 'real' sample measurements. For clarity, two examples of the described scenarios are presented:

Example 1 In an operationally defined procedure, such as the determination of malathion in oranges using extraction procedure A, the combination of the mean error with other uncertainty components is expected to produce confidence intervals overlapping the ones for the analysis of the same sample using extraction procedure B, even if the extraction procedures have significantly different efficiencies.

Example 2 If measurement procedure performance is dominated by the liquid/liquid extraction of the analyte, analyte losses are expected due to its partition in the two phases producing analyte recoveries below 100 %. However, if observed mean analyte recovery is above 100 %, the results of unknown samples should not be corrected for recovery since the positive error observed in reference material analysis has the chance of not occurring in the analysis of unknown samples. In this case, mean error is combined with other uncertainty components of measurements not corrected for the mean error.

The decision to correct or not to correct the measurement error or recovery that was observed in the analysis of a reference material, in the analysis results for the unknown materials, has an impact on measurement traceability that must be considered. Only if systematic effects observed in the analysis of a reference material are corrected in the measurement results of the unknown item, the results are traceable to the value embodied in the reference material. da Silva and Camões [13] discussed that taking the mean recovery in the uncertainty budget or to correct measurement results for observed recovery does not guarantee equivalent compliance decisions from the same measurement.

This work discusses the management of systematic effects by determining the recovery from the analysis of adequate reference materials and by correcting recovery if it is significantly different from 100 % taking the uncertainty of estimated recovery into account.

Barwick and Ellison [14, 15] developed strategies for evaluating mean recovery from the analysis of a certified reference material, samples without native analyte spiked at the same level, the same sample with native analyte spiked at the same or different levels, or a sample characterised by a reference procedure. However, these authors did not discuss how to assess mean recovery if at least two of these reference materials are used (e.g. recovery estimated from the analysis of two certified reference materials and ten samples with different levels of native analyte and spiked at different levels).

The Nordtest report for the evaluation of the measurement uncertainty [12] presents approximate algorithms for estimating trueness uncertainty from different reference materials, assuming some uncertainty components are negligible and the combination of measurement errors on different mathematical expressions allows for an approximate quantification of the impact of systematic effects on the measurement results. However, Nordtest approximations can be too optimistic or pessimistic depending on the relevant details of the trueness tests, such as the covered quantity levels and diversity of reference value uncertainties, suggesting the need for alternative approaches.

This work presents a methodology to assess mean recovery from the analysis of independent reference materials of different types. The method is based on the propagation of uncertainty components for models where the measurements precision varies with the quantity of interest and also considers the metrological significance of the mean measurement error. This methodology is applicable to cases where measurements of the native and of the spiked quantities are affected by relevant and significantly different uncertainties. This work extends methodologies proposed by Barwick and Ellison [14, 15] for evaluating the uncertainty associated with the observed mean recovery, to the determination of recovery from the analysis of a larger diversity of materials.

This methodology was successfully applied to the determination of metals in natural water by ICP-OES.

Theory

The theory is divided into two parts, i.e. the art of recovery evaluation and in the description of a novel methodology to assess mean recovery from a large diversity of reference materials. The impact of recovery test precision conditions on the assessment of systematic effects is discussed in detail.

Recovery estimation from one reference material

Barwick and Ellison [14, 15] proposed general algorithms for estimating mean recovery, $\bar{R}$, and the respective recovery uncertainty, $u_{\bar{R}}$, from the analysis of a reference material from two reference material types, i.e. a reference material external to the measurement procedure and a reference material internal to the measurement procedure.

Reference material external to the measurement procedure

If the reference material is prepared independently of measurements performed by the assessed measurement procedure, Eq. (1) is used to estimate $u_{\bar{R}}$:

$$u_{\bar{R}} = \bar{R}\sqrt{\left(\frac{s_R}{\bar{R}\sqrt{n}}\right)^2 + \left(\frac{u_Q}{Q}\right)^2} \quad (1)$$

where $\bar{R}$ is the mean recovery ($\bar{R} = \bar{q}/Q$; $\bar{q}$ and Q are the estimated mean and reference quantity values, respectively), s_R the standard deviation of estimated n recovery values and u_Q the standard uncertainty of Q. Usually, s_R is estimated under intermediate precision conditions to allow that $u_{\bar{R}}$ will be applicable to tests performed in subsequent days. This equation is applicable to recovery estimated from the analysis of a certified reference material, materials with negligible native quantity spiked at the same level of the quantity of interest and a material characterised by a reference procedure. In these cases, Q is the certified value, spiked value or value estimated by the reference procedure, respectively. This equation combines the standard uncertainty of $\bar{q}$ and Q using the law of propagation of uncertainty, where the relative standard uncertainty of $\bar{q}$ is equivalent to the relative standard deviation of the mean recovery ($s_R/(\bar{R}\sqrt{n})$).

All systematic effects affecting measurements, such as the ones resulting from the sample preparation, instrument calibration and matrix effects are combined in the estimated recovery. Equation 1 does not take into account the impact of measurement precision, typically the intermediate precision, in the measurement uncertainty since this component is to be accounted for by the measurement precision component.

Reference material internal to the measurement procedure

If the recovery is estimated from the analysis of a material with native quantity before and after spiking at a specific level of the quantity of interest, making recovery estimation dependent of native quantity determination, Eqs. (2) and (3) can be used to estimate mean recovery, $\bar{R}$, and the respective recovery standard uncertainty, $u_{\bar{R}}$.

$$\bar{R} = \frac{\bar{q} - \bar{q}_0}{q_+} \quad (2)$$

where $\bar{q}$ and $\bar{q}_0$ are the calculated mean values of the quantity of interest after and before spiking, respectively, and q_+ is the spiked quantity.

$$u_{\bar{R}} = \bar{R}\sqrt{\left(\frac{\frac{s^2(q)}{n} + \frac{s^2(q_0)}{m}}{(\bar{q} - \bar{q}_0)^2}\right) + \left(\frac{u(q_+)}{q_+}\right)^2} \quad (3)$$

where $s(q)$ and $s(q_0)$ are the standard deviations of estimated n and m replicate results of material analysis after and before spiking, and $u(q_+)$ the standard uncertainty of q_+.

In most cases, each pair of estimated quantities in the material after, q_i, and before, q_{0i}, spiking is determined under repeatability conditions (i.e. in a short period of time and using the same analyst and equipment combination), and $s(q)$ and $s(q_0)$ are the repeatability standard deviations. Equation (3) represents the combination of the uncertainty components of the variables in Eq. (2). In these cases, systematic effects quantification is affected by random effects observed under repeatability conditions. The assessed systematic effects can be divided into components that are constant and specific for the daily measurement runs, the laboratory and, if relevant, the measurement procedure. In operationally defined measurement procedures, the systematic effects attributed to the measurement procedure are, by definition, null [2]. The components of systematic effects attributed to the daily measurement run and to the laboratory are not cancelled in operationally defined measurements.

If the estimated quantities q_i and q_{0i} are determined on different days, the $s(q)$ and $s(q_0)$ are the intermediate precision standard deviations that quantify random effects responsible for the difference between $\bar{q}$ and $\bar{q}_0$. In these cases, the mean recovery assesses in particular the combination of systematic effects associated with the laboratory and, if relevant, the measurement procedure.

After $\bar{R}$ and $u_{\bar{R}}$ are estimated, it is tested whether $\bar{R}$ is significantly different from the ideal value of 1 by testing the following condition:

$$\frac{|1 - \bar{R}|}{u_{\bar{R}}} \leq t_v^{95\,\%} \quad (4)$$

where $t_v^{95\,\%}$ is the two-tailed Student's t for the degrees of freedom, v, of $u_{\bar{R}}$ and a 95 % confidence level. If the condition in Eq. (4) is true, the $\bar{R}$ is metrologically equivalent to 1 and no recovery correction of the original measurement results is required. If the condition in Eq. (4) is not true, a correction of the original measured quantity values of the unknown samples should be considered by multiplying the measured results by the reverse of the mean recovery ($1/\bar{R}$).

Barwick and Ellison [14, 15] also discussed how to estimate an additional uncertainty component for when recovery estimated for one quantity level/matrix combination is used to estimate measurement trueness for another quantity value/matrix combination. This approach relies on assessing measurement trueness from an adequate diversity of relevant effects affecting systematic effects, such as different matrixes of the measurement scope. If systematic

effects vary significantly with the analysed matrix, the standard deviation of the mean of recovery estimated for different matrices should be considered as an additional uncertainty component for trueness.

Recovery estimation from various reference materials

This section describes the algorithms developed and applied in this work.

Reference material external to the measurement procedure

If recovery is estimated from the analysis of N reference materials prepared independently of measurements performed by the assessed procedure and each reference material is analysed n_i times, the $\bar{R}$ is estimated by Eq. (5).

$$\bar{R} = \sum_{i=1}^{N} \left(\frac{\bar{q}_i}{Q_i} \right) \Big/ N \tag{5}$$

where $\bar{q}_i$ and Q_i are the estimated mean ($\bar{q}_i = \sum q_{ij}/n_i$, where q_{ij} is the jth replicate of reference material i analysis; $j = 1$ to n_i) and reference values of reference material i, respectively.

If the replicate analysis of the reference materials is performed on different days, since the procedure is to be used over an extended period of time, $u_{\bar{R}}$, is estimated by Eq. (6).

$$u_{\bar{R}} = \sqrt{\sum_{i=1}^{N} \left\{ \left(\frac{\bar{q}_i}{Q_i} \right)^2 \left[\left(\frac{s(q_i)}{\bar{q}_i \sqrt{n_i}} \right)^2 + \left(\frac{u(Q_i)}{Q_i} \right)^2 \right] \right\}} \Big/ N \tag{6}$$

where $s(q_i)$ is the intermediate precision standard deviation of q_{ij} values and $u(Q_i)$ the standard uncertainty of Q_i. If the reference materials have equivalent Q_i, the same $s(q_i)$ (e.g. a pooled intermediate precision standard deviation) can be considered. Models of intermediate precision variation with the quantity value can also be used to estimate $s(q_i)$, in particular if Q_i are significantly different [5, 6]. The estimated $\bar{R}$ is not focused on the quantification of systematic effects attributed to the daily run since it varies between runs. The intermediate precision standard deviation quantifies the combination of pure random effects with the variation of between run systematic effects. The repeatability standard deviation quantifies pure random effects.

If the reference materials are analysed under repeatability conditions, for instance when the measurement procedure is to be validated and used in a single day due to a request for urgent sample analysis, the $s(q_i)$ is the repeatability standard deviation and $\bar{R}$ assesses all possible systematic effects including the one attributed to the specific daily run.

Replicate analysis of the reference material should be performed in the same precision conditions (i.e. repeatability or intermediate precision conditions).

If each studied reference material is analysed once (i.e. $n_i = 1$), Eq. (6) is not converted into Eq. (1) since Q_i are assumed to be independent. Equation (6) is converted into Eq. (1) when only one reference material is analysed making $N = 1$.

Reference material internal to the procedure

If N materials with independent, different or equivalent, native quantity levels are spiked at independent levels, and materials are quantified n_i and m_i times after and before spiking, respectively ($i = 1$ to N), the mean recovery is estimated by Eq. (7).

$$\bar{R} = \sum_{i=1}^{N} \frac{\bar{q}_i - \bar{q}_{0i}}{q_{+i}} \Big/ N \tag{7}$$

where $\bar{q}_i$ and $\bar{q}_{0i}$ are the estimated mean quantities of material i after and before spiking, respectively, and q_{+i} the spiked quantity of material i ($\bar{q}_i = \sum q_{ij}/n_i$, where q_{ij} is the jth replicate result of the analysis of material i after spiking ($j = 1$ to n_i) and $\bar{q}_{0i} = \sum q_{0ik}/m_i$, where q_{0ik} is the kth replicate result of the analysis of material i before spiking ($k = 1$ to m_i)). If materials, before and after spiking, are analysed under repeatability conditions, the standard uncertainty, $u_{\bar{R}}$, of the mean recovery (Eq. (7)) is estimated by Eq. (8).

$$u_{\bar{R}} = \sqrt{\sum_{i=1}^{N} \left\{ \left(\frac{\bar{q}_i - \bar{q}_{0i}}{q_{+i}} \right)^2 \left[\frac{\frac{s^2(q_i)}{n_i} + \frac{s^2(q_{0i})}{m_i}}{(\bar{q}_i - \bar{q}_{0i})^2} + \left(\frac{u(q_{+i})}{q_{+i}} \right)^2 \right] \right\}} \Big/ N \tag{8}$$

where $s(q_i)$ and $s(q_{0i})$ are the repeatability standard deviations of q_{ij} and q_{0ik} replicate results, respectively, and $u(q_{+i})$ the standard uncertainty of q_{+i}. If n_i and m_i are smaller than 10, the $s(q_i)$ and $s(q_{0i})$ can be estimated from previously developed models of the variation of the standard deviation of the repeatability with the measured quantity associated with a larger number of degrees of freedom [5, 6]. Since precision conditions considered in Eq. (8) are repeatability conditions, the systematic effects assessed from estimated $\bar{R}$ and $u_{\bar{R}}$ are the ones observed within a run, in the laboratory and, for rational measurements, attributed to measurement procedure principles.

In the uncommon situations where materials after and before spiking are analysed on different days (i.e. under intermediate precision conditions), the $s(q_i)$ and $s(q_{0i})$ are intermediate precision standard deviations.

Equation (8) is not applicable to data collected under different precision conditions.

If each material after and before spiking is analysed once, Eq. (8) is simplified to Eq. (9):

$$u_{\bar{R}} = \sqrt{\sum_{i=1}^{N}\left\{\left(\frac{q_i - q_{0i}}{q_{+i}}\right)^2\left[\frac{s^2(q_i)+s^2(q_{0i})}{(q_i-q_{0i})^2}+\left(\frac{u(q_{+i})}{q_{+i}}\right)^2\right]\right\}}\Big/N \tag{9}$$

Reference material internal to the procedure: liquid reference materials internal to the procedure

In the analysis of liquid samples spiked with a standard solution volume, native quantity is diluted and, if relevant, this dilution should be taken into account in recovery assessment.

If N pairs of samples before and after spiking are analysed, the $\bar{R}$ is estimated by Eq. (10):

$$\begin{aligned}\bar{R} &= \sum_{i=1}^{N}\frac{\bar{\gamma}_i - \bar{\gamma}_{0i}\cdot[(V_{Ai}-V_{1i})/V_{Ai}]}{N\cdot\gamma_{Si}(V_{1i}/V_{Ai})}\\ &= \sum_{i=1}^{N}\frac{\bar{\gamma}_i\cdot V_{Ai} - \bar{\gamma}_{0i}\cdot(V_{Ai}-V_{1i})}{N\cdot\gamma_{Si}\cdot V_{1i}}\end{aligned} \tag{10}$$

where $\bar{\gamma}_i$ and $\bar{\gamma}_{0i}$ are the estimated mean mass concentrations of sample i after and before spiking.

The $u_{\bar{R}}$ is estimated by Eq. (11), which consists of the application of the law of propagation of uncertainty to combine standard uncertainties of Eq. (10) variables:

$$\begin{aligned}u_{\bar{R}} &= \sqrt{\sum_{i=1}^{N}\left\{\begin{array}{r}\left(\frac{\partial\bar{R}}{\partial\bar{\gamma}_i}\cdot u(\bar{\gamma}_i)\right)^2+\left(\frac{\partial\bar{R}}{\partial\bar{\gamma}_{0i}}\cdot u(\bar{\gamma}_{0i})\right)^2+\left(\frac{\partial\bar{R}}{\partial V_{Ai}}\cdot u(V_{Ai})\right)^2+\\ \left(\frac{\partial\bar{R}}{\partial V_{1i}}\cdot u(V_{1i})\right)^2+\left(\frac{\partial\bar{R}}{\partial\gamma_{Si}}\cdot u(\gamma_{Si})\right)^2\end{array}\right\}}\\ &= \sqrt{\sum_{i=1}^{N}\left\{\begin{array}{r}\left(\frac{V_{Ai}}{N\cdot\gamma_{Si}\cdot V_{1i}}\cdot\frac{s(\gamma_i)}{\sqrt{n_i}}\right)^2+\left(\frac{V_{1i}-V_{Ai}}{N\cdot\gamma_{Si}\cdot V_{1i}}\cdot\frac{s(\gamma_{0i})}{\sqrt{m_i}}\right)^2+\left(\frac{\bar{\gamma}_i-\bar{\gamma}_{0i}}{N\cdot\gamma_{Si}\cdot V_{1i}}\cdot u(V_{Ai})\right)^2+\\ \left(\frac{\bar{\gamma}_{0i}-(\gamma_{Si}\cdot\bar{R}_i)}{N\cdot\gamma_{Si}\cdot V_{1i}}\cdot u(V_{1i})\right)^2+\left(-\frac{\bar{R}_i}{N\cdot\gamma_{Si}}\cdot u(\gamma_{Si})\right)^2\end{array}\right\}}\end{aligned} \tag{11}$$

The most convenient way to perform these spikes is by taking a volumetric flask with volume, V_A, adding spiked volume, V_1, of the standard solution and filling up the flask with the sample solution. In this case, the spiked mass concentration of solution i, γ_{+i}, is $[\gamma_{+i} = \gamma_{Si}(V_{1i}/V_{Ai})]$ where γ_{Si} is the mass concentration of the standard solution. (The notation q is changed to γ since the gamma is the notation indicated for mass concentrations.) The native quantity in spiked sample i, $\gamma_{0(d)i}$, is $\gamma_{0(d)i} = \gamma_{0i}[(V_{Ai}-V_{1i})/V_{Ai}]$, where γ_{0i} is the native mass concentration. The native sample dilution factor in spiked samples (i.e. $[(V_{Ai}-V_{1i})/V_{Ai}]$) should not be smaller than 80 % to guarantee that the recovery in the diluted matrix will be representative of the recovery observed in undiluted samples. Even if strong matrix effects affect measurements, the dilution of about 20 % of the matrix should not produce matrix effects significantly different from those observed in undiluted matrices.

where $s(\gamma_i)$ and $s(\gamma_{0i})$ are the repeatability standard deviations of γ_{ij} ($j = 1$ to n_i) and γ_{0ik} ($k = 1$ to m_i) measurements if, for each recovery test i, measurements are performed under repeatability conditions and $\bar{R}_i$ is the mean recovery estimated from test i. The independent recovery tests (e.g. recovery tests $i = 1$ and $i = 2$) can be performed on the same or different days since this is irrelevant for Eq. (11).

If a volume, V_{1i}, of the stock solution of the quantity of interest (stock solution mass concentration γ_{Si}) is not the only one added to the flask (flask volume V_{Ai}) where the sample will be diluted, but $(p-1)$ additional volumes V_{2i} to V_{pi} of other solutions of the same solvent are being added with no relevant levels of the quantity of interest, recovery is estimated by Eq. (12). The additional solutions can be spikes of other analytes.

$$\bar{R} = \sum_{i=1}^{N} \frac{\bar{\gamma}_i - \bar{\gamma}_{0i} \cdot [(V_{Ai} - V_{1i} - V_{Bi})/V_{Ai}]}{N \cdot \gamma_{Si}(V_{1i}/V_{Ai})} = \sum_{i=1}^{N} \frac{\bar{\gamma}_i \cdot V_{Ai} - \bar{\gamma}_{0i} \cdot (V_{Ai} - V_{1i} - V_{Bi})}{N \cdot \gamma_{Si} \cdot V_{1i}} \quad (12)$$

where V_{Bi} is the sum of solution volumes, other than V_{1i}, added to diluted sample flask (i.e. $V_{Bi} = \sum_{i=2}^{p} V_{pi}$). The standard uncertainty of $\bar{R}$, determined by Eq. (12), is estimated by Eq. (13):

$$u_{\bar{R}} = \sqrt{\sum_{i=1}^{N} \left\{ \begin{array}{l} \left(\frac{\partial \bar{R}}{\partial \bar{\gamma}_i} \cdot u(\bar{\gamma}_i)\right)^2 + \left(\frac{\partial \bar{R}}{\partial \bar{\gamma}_{0i}} \cdot u(\bar{\gamma}_{0i})\right)^2 + \left(\frac{\partial \bar{R}}{\partial V_{Ai}} \cdot u(V_{Ai})\right)^2 + \\ \left(\frac{\partial \bar{R}}{\partial V_{1i}} \cdot u(V_{1i})\right)^2 + \left(\frac{\partial \bar{R}}{\partial \gamma_{Si}} \cdot u(\gamma_{Si})\right)^2 + \left(\frac{\partial \bar{R}}{\partial V_{Bi}} \cdot u(V_{Bi})\right)^2 \end{array} \right\}}$$
$$= \sqrt{\sum_{i=1}^{N} \left\{ \begin{array}{l} \left(\frac{V_{Ai}}{N \cdot \gamma_{Si} \cdot V_{1i}} \cdot \frac{s(\gamma_i)}{\sqrt{n_i}}\right)^2 + \left(\frac{V_{1i} - V_{Ai}}{N \cdot \gamma_{Si} \cdot V_{1i}} \cdot \frac{s(\gamma_{0i})}{\sqrt{m_i}}\right)^2 + \left(\frac{\bar{\gamma}_i - \bar{\gamma}_{0i}}{N \cdot \gamma_{Si} \cdot V_{1i}} \cdot u(V_{Ai})\right)^2 + \\ \left(\frac{\bar{\gamma}_{0i} - (\gamma_{Si} \cdot \bar{R}_i)}{N \cdot \gamma_{Si} \cdot V_{1i}} \cdot u(V_{1i})\right)^2 + \left(-\frac{\bar{R}_i}{N \cdot \gamma_{Si}} \cdot u(\gamma_{Si})\right)^2 + \left(\frac{\bar{\gamma}_{0i}}{N \cdot \gamma_{Si} \cdot V_{1i}} \cdot u(V_{Bi})\right)^2 \end{array} \right\}} \quad (13)$$

Estimation of the precision of the recovery tests

Depending on the precision conditions affecting the estimated recovery, the mean recovery standard uncertainty can be determined using repeatability or intermediate precision standard deviations. The precision conditions affecting mean recovery will also determine which systematic effects are assessed from the mean recovery as discussed previously.

For the trueness test, reference materials external or internal to the measurement procedure can be analysed. For the case where these reference materials are analysed once or from a small number of tests, it is convenient to use prior models of precision variation with the quantity of interest build from an adequately large number of experimental data. For most analytical applications, precision estimation is adequate if it is associated with at least 14 degrees of freedom [16].

Ideally, the models of precision variation with the quantity should be based on information collected at several quantity levels. However, the information from a single level can be used to model precision in a wide range if some general trends in measurement precision are considered.

In most classical and instrumental measurements, precision standard deviation is approximately constant in a narrow range and tends to increase as the quantity increases in a wider range. On the other hand, the precision relative

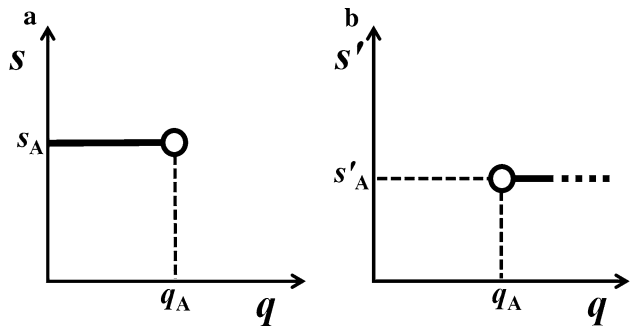

Fig. 1 Model of measurement precision variation with the quantity of interest built from precision estimated at one quantity, q_A. **a** The precision standard deviation, s_A, estimated at a specific quantity, q_A, overestimates precision below q_A; **b** the precision relative standard deviation, s'_A ($s'_A = s_A/q_A$), estimated at q_A, overestimates precision above q_A

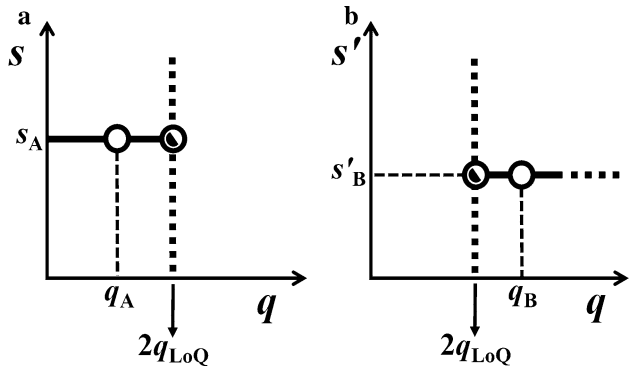

Fig. 2 Model of measurement precision variation with the quantity of interest (thicker line) built from precision estimated at quantities q_A and q_B positioned below and above two times the Limit of Quantification ($2\gamma_{LoQ}$), respectively. The s_A and s'_B represent the absolute and relative precision standard deviations associated with q_A and q_B, respectively

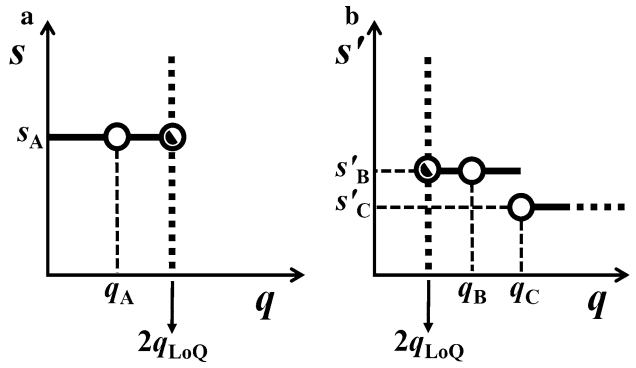

Fig. 3 Model of measurement precision variation with the quantity of interest built from precision estimated at quantities q_A, q_B and q_C. q_A is smaller than two times the Limit of Quantification ($2\gamma_{LoQ}$), and q_B and q_C are larger than $2\gamma_{LoQ}$. The s_A, s'_B and s'_C represent the absolute and relative precision standard deviations associated with q_A, q_B and q_C, respectively (the apostrophes identify relative quantities)

standard deviation tends to decrease in an abrupt way from the Limit of Detection to two times the Limit of Quantification ($2q_{LoQ}$), decreasing slightly after this level [5, 6]. Therefore, regardless of the level of the quantity of interest at which precision was estimated, it can be assumed, by approximation, that the observed precision standard deviation overestimates precision below the studied level and the precision relative standard deviation overestimates precision above the studied level (Fig. 1a, b) [5, 6]. If precision is estimated at various levels, adequate step models can be built. Figures 2a, b and 3a, b present examples where precision models are defined from precision estimated at two or three levels positioned below and above $2q_{LoQ}$. If more levels are studied, more complex models, such as a linear relation between the precision standard deviation and the quantity of interest, can be built [9, 17, 18].

Experimental

The developed methodology for mean recovery assessment, particularly the presented algorithms to estimate recovery from the analysis of different liquid samples with native quantity before and after spiking, was applied to the determination of dissolved metals in natural waters by inductively coupled plasma atomic emission spectrometry (ICP-OES). Samples were analysed after filtration with a 0.45 µm pore cellulose acetate filter. The top-down approach based on in-house validation data was used to estimate measurement uncertainty by combining two major uncertainty components: recovery/trueness and precision components. No relevant additional uncertainty components were identified.

Material

The volumetric operations were performed using class A volumetric glassware subject to an adequate washing procedure. The 0.45 µm pore cellulose acetate filter was purchased from Pall Corporation (New York, USA). A Thermo Scientific (Waltham, MA, USA) iCAP 7400 ICP-OES Duo spectrometer was used in quantifications.

Chemicals

Purified chemicals adequate for performed analysis and checked with blank tests were used. Single-element stock solutions purchased from Merck (Darmstadt, Germany) with a reference value of (1000 ± 10) mg L^{-1} (coverage factor of 2) of the metal were used. Different lots of Merck solutions were used to prepare calibrators and to spike samples to guarantee deviation in stock solution values do not cancel in recovery tests. Merck solutions have metal contents traceable to the unit mg L^{-1} of the International System of Units (SI) checked through the analysis of the corresponding Standard Reference Materials (NIST SRM®) produced by the National Institute of Standards and Technology of the USA. Suprapur grade nitric acid purchased from Merck (Darmstadt, Germany) was used.

Analysed samples

Samples of surface natural waters, collected in rivers and bayous, were spiked and analysed to estimate recovery.

Proficiency test

The developed methodology was assessed through the participation in two proficiency tests: (1) Aquacheck 1S, Round 485, May 2015—Soft water—Major Inorganic Components [19]; (2) RELACRE EAA, 1st Round, June 2015, Drinking water [20].

Measurement procedure

The measurement procedure involves sample filtration and acidification to 0.2 % nitric acid with a negligible volume and, if relevant, dilution before collecting ICP-OES signals. The spectrometer is subject to an analytical calibration before samples analysis. Table 1 lists the studied elements, the wavelength of emission lines, the plasma view configurations and the calibration range. For some elements, instrument response was calibrated in a low and a high mass concentration range to allow direct measurement of samples with higher concentrations. The details of calibrators preparation are omitted for simplicity.

Table 1 List of elements analysed in natural water by ICP-OES, relevant instrument details and studied calibration ranges

Element	Emission wavelength (nm)	Plasma view configuration	Calibration range[a] (mg L^{-1})
Na	589.592	Radial	0.5–5
	589.592	Radial	5–50
K	766.490	Radial	0.4–2
	766.490	Radial	2–20
Mg	285.210	Radial	0.2–2
Ca	317.933	Radial	0.2–2
	315.887	Radial	2–6
Cr	267.716	Axial	0.002–0.02
Mn	257.610	Axial	0.004–0.02
	257.610	Axial	0.02–0.2
Fe	259.940	Axial	0.01–0.1
	259.940	Radial	0.1–1
Cu	324.754	Axial	0.004–0.02

[a] The calibration range presents the lower and higher mass concentrations of quantitative calibrators (i.e. excluding the blank considered in the calibration)

Calibrations were performed at six levels, including blank, with approximately equidistant mass concentrations.

Results and discussion

Measurement performance assessment

The following sections describe how different performance parameters were determined and the maximum values for these parameters. Since the decision about measurement procedure fitness for the intended use is based in the comparison of the Limit of Quantification, γ_{LoQ}, and uncertainty with the Maximum Limit of Quantification, γ_{LoQ}^{Max}, and the target uncertainty, respectively, the maximum values for measurements repeatability and intermediate precision are only indicative [6]. The γ_{LoQ} is also relevant to set models of the variation of measurement precision throughout the calibration range (see sections "Results and discussion—Measurement performance assessment—Measurement repeatability" and "Measurement intermediate precision").

The linearity of the variation of the ICP-OES emission with the mass concentration of the studied element in analysed solution was assessed, and a linear regression model was used to build the calibration curve. Relevant deviations from linearity can make the analyte recovery observed by interpolating signal in one portion of the calibration curve not applicable to interpolations performed in another segment of the calibration curve.

Limit of Quantification

The assessment of measurement procedure performance started with γ_{LoQ} determination ($\gamma_{LoQ} = 10s(\gamma_{CS})$), by taking ten times the standard deviation, $s(\gamma_{CS})$, of at least ten ($n \geq 10$) measurement results, γ_{CSi} ($i = 1$ to n), obtained on different days, of a control standard with a quantity level, Γ_{CS}, equivalent to the expected γ_{LoQ}. If the estimated γ_{LoQ} is more than five times different from Γ_{CS} (i.e. if $(\bar{\gamma}_{CS}/\Gamma_{CS}) < 0.2$ or $(\bar{\gamma}_{CS}/\Gamma_{CS}) > 5$), a control standard with a different concentration should be prepared and analysed on different days to guarantee $s(\gamma_{CS})$ adequately estimates the precision at the γ_{LoQ}. The trueness of control standard measurements was assessed, in a pragmatic way, by checking whether the absolute value of the difference $(\bar{\gamma}_{CS} - \Gamma_{CS})$ is smaller than the standard deviation, $s(\bar{\gamma}_{CS})$, of $\bar{\gamma}_{CS}$ times the Student's t for $(n - 1)$ degrees of freedom and 99 % confidence level, t ($|\bar{\gamma}_{CS} - \Gamma_{CS}| \leq t \cdot s(\gamma_{CSi})/\sqrt{n}$), where $\bar{\gamma}_{CS}$ is the mean of γ_{CSi} values, $\bar{\gamma}_{CS} = \sum \gamma_{CSi}/n$, and $s(\bar{\gamma}_{CS}) = s(\gamma_{CSi})/\sqrt{n}$. If this condition is valid, no relevant systematic effects affect quantifications at the γ_{LoQ}. This condition is not adequate to compare $\bar{\gamma}_{CS}$ with Γ_{CS} if Γ_{CS} is associated with a relevant uncertainty.

The γ_{LoQ}^{Max} is 30 % of the 'environmental quality standard' value set by the national regulator for water status monitoring as defined in Directive 2009/90/EC [21]. If no reference for the environmental monitoring is set, the γ_{LoQ}^{Max} is defined from the maximum Limit of Detection, γ_{LoD}^{Max}, set for the analysis of drinking water in Council Directive

Table 2 Calibration range, and target and observed performance parameters

Element	Calib. range (mg L^{-1})	γ_{LoQ}^{Max} (mg L^{-1})	γ_{LoQ} (mg L^{-1})	$s_{r(I)}$ (mg L^{-1})	$s_{r(I)}/\gamma_{LoQ}$(%)	$s'_{r(II)}$ (%)	$s_{IP(I)}$[g] (mg L^{-1})	$s'_{IP(II)}$ (%)	$\bar{R}$[h] (%)	$u_{\bar{R}}$ (%)	Target uncertainty[i]	Relative expanded uncertainty (%)[k]
Na	0.5–5	67[a]	0.24	0.0075[e]	3.1	1.5	0.024	4.8	98.6	1.7	26 mg L^{-1}	10.3
Na	5–50	[b]	5.2	0.083	1.6	1.3	0.52	2.0	99.50	0.88	26 mg L^{-1}	20.9–4.5
K	0.4–2	[c]	0.27	0.014[e]	5.2	2.5	0.027	5.2	103.0	1.7	50 %[j]	14.0–10.9
K	2–20	[c]	2.2	0.060	2.7	0.75	0.22	2.2	98.95	0.63	50 %[j]	22.0–4.5
Mg	0.2–2	[c]	0.23	0.0048	2.1	1.2	0.023	4.1	97.2	1.2	50 %[j]	23.4–8.6
Ca	0.2–2	[c]	0.13	0.0054[e]	4.2	2.0	0.013	3.2	99.0	1.7	50 %[j]	13.6–7.4
Ca	2–6	[c]	< 2	–[f]	–	1.4	–[f]	1.5	100.6	1.3	50 %[j]	3.9
Cr	0.002–0.02	0.0014[d] 0.017[a]	0.0014	6.4×10^{-5}[e]	4.6	2.2	1.4×10^{-4}	3.3	103.8	1.6	0.0064 mg L^{-1}	14.5–7.2
Mn	0.004–0.02	0.017[a]	0.0024	4.3×10^{-5}[e]	1.8	0.90	2.4×10^{-4}	2.7	98.5	1.3	0.0064 mg L^{-1}	12.3–6.0
Mn	0.02–0.2	[b]	0.029	7.0×10^{-4}	2.4	0.58	0.0029	2.9	100.4	1.1	0.0064 mg L^{-1}	28.9–6.2
Fe	0.01–0.1	0.067[a]	0.0098	2.6×10^{-4}	2.7	2.4	9.8×10^{-4}	4.8	101.8	1.4	0.026 mg L^{-1}	19.7–9.9
Fe	0.1–1	[b]	< 0.1	–[f]	–	1.2	0.0026	2.6	97.86	0.83	0.026 mg L^{-1}	5.5
Cu	0.004–0.02	0.0023[d] 0.67[a]	0.0029	1.4×10^{-4}[e]	4.8	2.5	2.9×10^{-4}	4.6	101.5	1.8	0.26 mg L^{-1}	14.9–9.8

γ_{LoQ}^{Max}: Maximum Limit of Quantification; $s_{r(I)}$ and $s'_{r(II)}$: Absolute and relative repeatability standard deviations in intervals I and II, respectively (interval I: between γ_{LoQ} and $2\gamma_{LoQ}$, inclusive; interval II: larger than $2\gamma_{LoQ}$; $s_{IP(I)}$ and $s'_{IP(II)}$: Absolute and relative intermediate precision standard deviation in intervals I and II, respectively; $\bar{R}$ and $u_{\bar{R}}$: Mean recovery and respective standard uncertainty

[a] γ_{LoQ}^{Max} estimated as ten-thirds the maximum limit of detection set in Directive 98/83/EC [22] for drinking water monitoring

[b] No target γ_{LoQ} is defined for the higher calibration range of each element analysis

[c] No target value set due to low toxicological relevance

[d] γ_{LoQ}^{Max} estimated as 30 % of the 'environmental quality standard' according to Directive 2009/90/EC [21] where the quality standard is defined in the 'Portuguese Hydrographic Region Management Plan for 2016–2021' [23]

[e] Estimated indirectly as $(s'_{r(II)} \cdot 2 \cdot \gamma_{LoQ})$ since not more than six duplicates in concentration interval I (i.e. between γ_{LoQ} and $2\gamma_{LoQ}$, inclusive) were collected

[f] Not estimated since quantitative calibrators have a mass concentration larger than $2\gamma_{LoQ}$

[g] The $s_{IP(I)}/\gamma_{LoQ}$ is 10 % since $s_{IP(I)}$ is used for the determination of γ_{LoQ}

[h] Not relevant systematic effects for 95 % confidence level except for Mg, Cr and Fe in interval I where systematic effects are only negligible for 99 % confidence level

[i] Absolute values (mg L^{-1}) as defined by combining performance parameters set in Directive 98/83/EC [22] using criteria proposed in the Eurachem/CITAC guide for setting the target uncertainty [6]

[j] The relative target uncertainty defined in Directive 2009/90/EC [21]

[k] Uncertainty expanded to approximately 95 % confidence level using coverage factor of 2

98/83/EC [22], assuming that the γ_{LoQ}^{Max} is 10/3 larger than the γ_{LoD}^{Max}. For elements with no limits set due to its low toxicological relevance, the γ_{LoQ} should be smaller than analysed sample concentrations.

Table 2 presents the defined γ_{LoQ}^{Max} and the estimated γ_{LoQ}. Since quantifications of Γ_{CS} are not affected by relevant systematic effects and γ_{LoQ} is not significantly larger than γ_{LoQ}^{Max}, measurement procedure γ_{LoQ} is fit for the intended use. For measurements of the mass concentration of Cu, the estimated γ_{LoQ} (i.e. 0.0029 mg L^{-1}) is not significantly larger than the γ_{LoQ}^{Max} set from 'environmental quality standards' (i.e. 0.0023 mg L^{-1}) taking the expected variability of precision estimates [5, 6]. The calibration ranges for which no γ_{LoQ}^{Max} is set, have a γ_{LoQ} or lower calibration level, excluding the blank, smaller than levels in studied samples.

Instrument signal linearity

The linearity of instrument response variation with the analyte mass concentration was tested with the ANOVA lack-of-fit test (ANOVA-LOF) [24] or the Chi-squared lack-of-fit test (χ^2-test) [25] applicable to calibration ranges

where signal variances are constant or vary with the quantity of interest, respectively. The homogeneity of signal variance was tested with Levene's test [24]. If instrument signal varies linearly with the concentration, the least squares regression model, LSRM, is adequate to estimate the intercept and the slope of the calibration curve regardless of the homogeneity or heterogeneity of signal's variance [10]. da Silva [26] presented experimental evidences of the statistical equivalence of results estimated by the linear unweighted (i.e. the LSRM) or the linear weighted regression model even if signal variance varies in the calibration range.

The instrument signal varies linearly in the studied mass concentration ranges of the various elements.

Measurement repeatability

The measurement repeatability was estimated from duplicate measurements, obtained under repeatability conditions, of different real samples. The range (the range is the absolute value of the difference), A_j, of duplicate results of N samples ($j = 1$ to N) with a mean value between the γ_{LoQ} and $2\gamma_{\mathrm{LoQ}}$ inclusive, designated interval I, were combined in the same mean range, $\bar{A}$ ($\bar{A} = \sum A_j / N$) to estimate the repeatability standard deviation, $s_{r(\mathrm{I})}$, in this concentration interval: $s_{r(\mathrm{I})} = \bar{A}/1.128$ [27]. For sample concentrations larger than the two times the γ_{LoQ} (i.e. in interval II), duplicate results relative ranges, A'_j (i.e. the range divided by the mean value), are combined in the same mean relative range, $\bar{A}'$, to estimate the relative repeatability standard deviation, $s'_{r(\mathrm{II})}$, in interval II: $s'_{r(\mathrm{II})} = \bar{A}'/1.128$. Table 2 presents estimated $s_{r(\mathrm{I})}$ and $s'_{r(\mathrm{II})}$. For the cases where less than six samples were analysed in 'interval I', the $s_{r(\mathrm{I})}$ is estimated indirectly using the $s'_{r(\mathrm{II})}$ ($s_{r(\mathrm{I})} = s'_{r(\mathrm{II})} \cdot 2 \cdot \gamma_{\mathrm{LoQ}}$).

The measurements repeatability was assumed to be fit for the intended use if the repeatability relative standard deviation is not larger than 5 % in interval I and not larger than 2.5 % in interval II. The $s_{r(\mathrm{I})}$ is compared with the target value after dividing it by the γ_{LoQ} to estimate the largest relative standard deviation in interval I. The $(s_{r(\mathrm{I})}/\gamma_{\mathrm{LoQ}})$ and $s'_{r(\mathrm{II})}$ are not significantly larger than 5 % and 2.5 %, respectively, proving repeatability is fit for the intended use. In K measurements between 0.4 mg L^{-1} and 2 mg L^{-1}, the $(s_{r(\mathrm{I})}/\gamma_{\mathrm{LoQ}})$ is only slightly larger than 5 % (i.e. 5.2 %).

Measurement intermediate precision

Intermediate precision of the measurements was estimated at two concentration levels from the analysis of control standards with values equivalent to the γ_{LoQ} and above $2\gamma_{\mathrm{LoQ}}$ (approximately in the middle of the calibration range). The intermediate precision standard deviation, $s_{\mathrm{IP(I)}}$, estimated at the γ_{LoQ}, is used to determine precision in concentration interval I (i.e. between γ_{LoQ} and $2\gamma_{\mathrm{LoQ}}$, inclusive) and the relative standard deviation of the second control standard results, $s'_{\mathrm{IP(II)}}$, used to estimate the relative precision in interval II (i.e. above $2\gamma_{\mathrm{LoQ}}$). Control standards are prepared independently of calibrators. Table 2 presents the estimated $s_{\mathrm{IP(I)}}$ and $s'_{\mathrm{IP(II)}}$. Intermediate precision of the measurements is considered fit for the intended use since $(s_{\mathrm{IP(I)}}/\gamma_{\mathrm{LoQ}})$ and $s'_{\mathrm{IP(II)}}$ are not significantly larger than 10 % and 5 %, respectively. The $s_{\mathrm{IP(I)}}/\gamma_{\mathrm{LoQ}}$ is exactly 10 % since $s_{\mathrm{IP(I)}}$ is used to estimate the γ_{LoQ} ($\gamma_{\mathrm{LoQ}} = 10 s_{\mathrm{IP(I)}}$). In the first calibration range of the determination of the mass concentration of K, the $s'_{\mathrm{IP(II)}}$ is slightly above 5 % (i.e. $s'_{\mathrm{IP(II)}} = 5.2$ %).

Measurement recovery

The measurement recovery was assessed from the analysis of real samples before and after spiking. The $\bar{R}$ and $u_{\bar{R}}$ were estimated as described previously in section "Reference material internal to the procedure: Liquid reference materials internal to the procedure" for all the calibration ranges. No target values for this uncertainty component alone are defined. Table 2 presents the estimated recovery and respective uncertainty. In all studied calibration ranges, except for the determination of Cr and Mg, and in the larger calibration range for Fe determination, estimated mean recovery is metrologically equivalent to 1 for a confidence level of 95 %. For the three specified cases, recovery becomes equivalent to 1 if a 99 % confidence level is considered. The estimated $u_{\bar{R}}$ are associated with a large number of degrees of freedom since more than 14 recovery tests were pooled. Therefore, it was decided not to correct results for recovery in all cases.

Measurement uncertainty

The precision and trueness uncertainty components were combined as relative standard uncertainties to estimate a combined standard uncertainty, u_{c}. The u_{c} was multiplied by a coverage factor of 2 to estimate the expanded uncertainty U_{c} for a confidence level of approximately 95 %. The large number of data used to estimate both uncertainty components guaranteed that this coverage factor is adequate. The degrees of freedom of the precision component are the degrees of freedom of the standard deviation of the intermediate precision. The degrees of freedom of the trueness component can be estimated by the Welch–Satterthwaite equation but should not be smaller than the degrees of freedom of the standard deviation of estimated recoveries [28, 29]. When two uncertainty components with equivalent impact on the model equation are

combined and components are associated with a similar number of degrees of freedom, the combined uncertainty has a number of degrees of freedom equivalent to the components' one.

Although the Commission Directive 2009/90/EC [21] for monitoring water status sets a maximum relative expanded uncertainty of 50 %, it was decided to apply some stricter performance criteria set for drinking water monitoring [22] for elements where a maximum permissible value is set for drinking water. The Directive 98/83/EC [22] defines maximum values for the intermediate precision standard deviation and for the mean error that were converted in a target uncertainty using the algorithm proposed in section 5.1.3 of the Eurachem/CITAC guide for setting the target measurement uncertainty [6]. The defined target uncertainties are smaller or equal to the proposed in Commission Directive 2009/90/EC [21].

Table 2 presents the expanded relative uncertainty of measurements performed in the various calibration ranges applicable to the analysis of samples requiring no dilution or a dilution with a negligible uncertainty. If the sample is diluted once, where the initial volume is not smaller than 0.5 mL and the final volume is not smaller than 5 mL measured using class A volumetric glassware, the dilution factor relative standard uncertainty is not larger than 1.2 % [26]. This dilution factor uncertainty is negligible if ICP-OES quantifications are associated with an expanded relative uncertainty not smaller than 7.2 % (7.2 % = 1.2 %·3·2, where factor 3 is used to increase the uncertainty to a significantly larger value and 2 to expand the standard uncertainty to a 95 % confidence level).

In all the quantifications performed in the studied ranges, except in two cases, the estimated uncertainty is smaller than the target uncertainty presented in Table 2. The determination of Mn in the calibration range with larger concentrations has an expanded uncertainty smaller than the target values below 0.096 mg L^{-1}. However, the maximum permissible manganese mass concentration in drinking water (i.e. 0.05 mg L^{-1}) suggests that the defined target uncertainty for the second half of the calibration range is too low. Similarly, the determination of Fe in the calibration range with lower concentrations is only associated with an expanded uncertainty smaller than the target uncertainty for quantified concentrations below 0.475 mg L^{-1}, where the maximum permissible mass concentration of iron in drinking water (i.e. 0.2 mg L^{-1}) is positioned. Therefore, the deviation to the initially defined target uncertainties is not critical in both these cases.

Figures 4, 5 and 6 present the developed models of relative expanded uncertainty variation with analyte concentration in the calibration curve. The figures also present the target measurement uncertainty. The axes of Figs. 4, 5 and 6 are logarithmic to allow representing significantly different ranges in the same graph. The logarithmic scale is rather convenient since many lines become straight lines.

Measurement traceability

Since quantifications are supported in calibrators prepared from Merck stock solutions using volumetric equipment that measure volume traceable to the SI unit metre, and no relevant lack of linearity or selectivity of ICP-OES response was observed, it can be concluded that the produced measurement results of unknown samples are directly traceable to the value embodied in the Merck stock solution and indirectly to the SI unit mg L^{-1}.

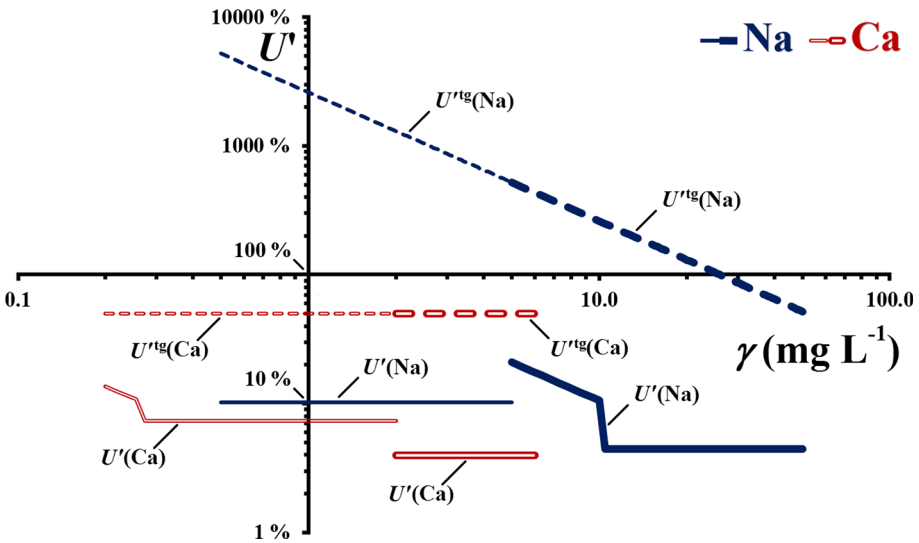

Fig. 4 Variation of estimated (U'; Na: —; Ca: =) and target (U'^{tg}; Na: - -;Ca: = =) relative expanded uncertainty, reported in percentage, with the measured mass concentrations of Na and Ca in water, γ (mg L^{-1}), in two calibration ranges. Continuous and dashed lines represent the estimated and target uncertainty, respectively. Calibration ranges with larger concentrations are represented by a thicker line. Both U' and γ values are presented on a logarithmic scale

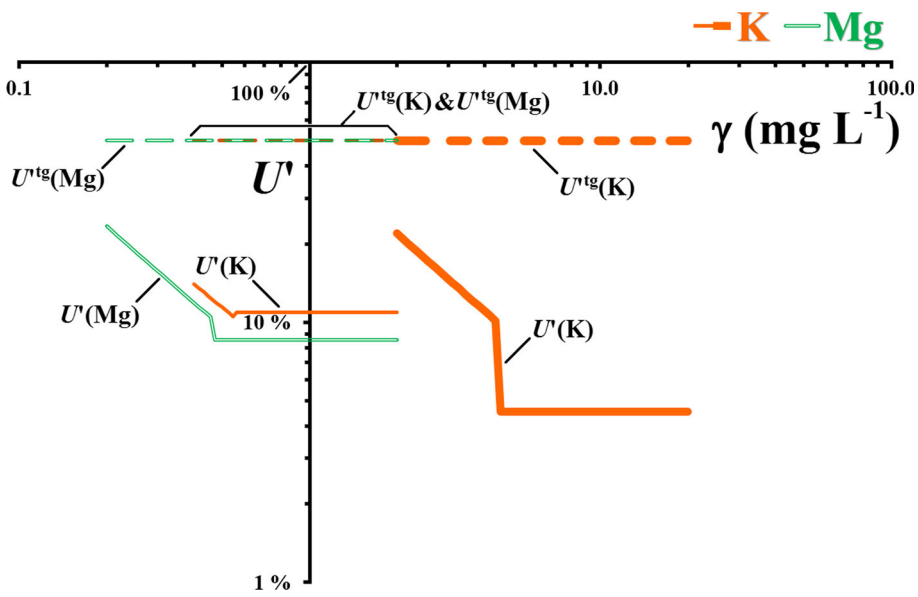

Fig. 5 Variation of estimated (U'; K: —; Mg: =) and target (U'^{tg}; K: - -;Mg: = =) relative expanded uncertainty, reported in percentage, with the measured mass concentrations of K and Mg in water, γ (mg L^{-1}), in one or two calibrations ranges. Continuous and dashed lines represent the estimated and target uncertainty, respectively. The calibration range for the determination of *K* with larger concentrations is represented by a thicker line. Both U' and γ values are presented on a logarithmic scale

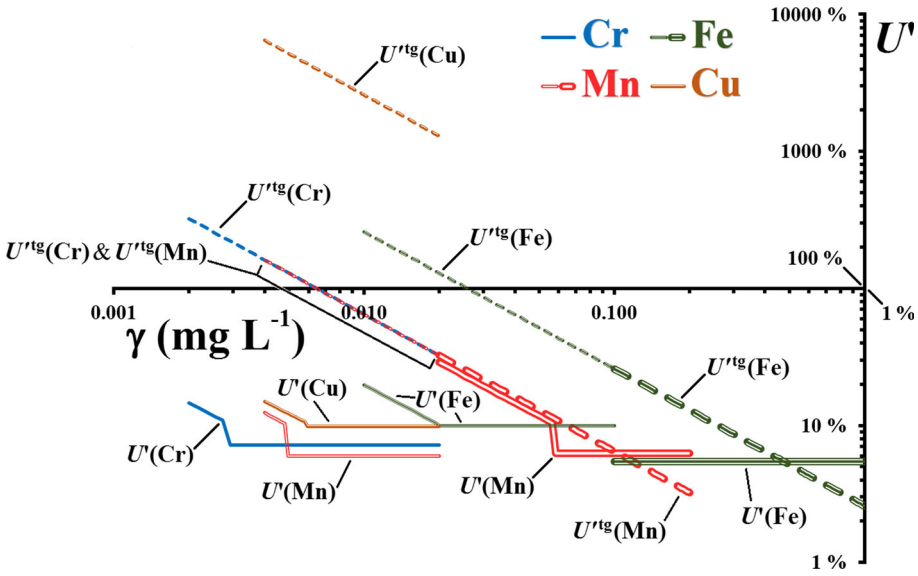

Fig. 6 Variation of estimated (U'; Cr: —; Fe: ≡ ; Mn: =, Cu: ═) and target (U'^{tg}; Cr: - -; Fe: ≡ ≡ ; Mn: = = , Cu: ═) relative expanded uncertainty, reported in percentage, with the measured mass concentrations of Cr, Fe, Mn and Cu in water, γ (mg L^{-1}), in one (Cr and Cu) or two (Fe and Mn) calibrations ranges. Continuous and dashed lines represent the estimated and target uncertainty, respectively. The calibration ranges for the determinations of Fe and Mn with larger concentrations are represented by a thicker line. Both U' and γ values are presented on a logarithmic scale

Proficiency test results

The developed measurement models were applied to the analysis of metals in water samples from two proficiency tests.

Table 3 presents the reference and estimated mass concentrations of metals in proficiency test samples and respective scores, namely the *z*-score and the E_n number [30]. The *z*-score is the ratio between the measurement error and half the maximum admissible error defined by the proficient test provider, being satisfactory between − 2 and 2. The E_n number is the ratio between the measurement error and the expanded uncertainty of the error assuming the reference and estimated values' uncertainties were expanded to a 95 % confidence level using a coverage factor of 2. Therefore, E_n numbers are satisfactory if have a value between − 1 and 1.

The results of Table 3 prove that the measurement error is within the acceptable range defined by the proficiency test providers and that the measurement uncertainty adequately predicts the measurement error. The reported expanded uncertainties are smaller than the target measurement uncertainty presented in Table 2.

The diversity of elements, calibration ranges and experimental data used in the uncertainty evaluations suggest that the developed algorithms and models are adequate to estimate measurement uncertainty. Although in some cases, the reported uncertainty is smaller than the one

Table 3 Results and scores of the participation in Aquacheck and RELACRE proficiency tests

Proficiency test	Element	Reference value (mg L^{-1})[c]	Estimated value (mg L^{-1})[c]	Calibration range (mg L^{-1})	*z*-score	E_n number
Aquacheck[a]	Na	10.110 ± 0.060	10.0 ± 1.0	0.5–5	− 0.15	− 0.10
	K	1.053 ± 0.013	1.05 ± 0.11	0.4–2	− 0.01	− 0.02
	Mg	2.350 ± 0.020	2.25 ± 0.40	0.2–2	− 0.40	− 0.25
	Ca	13.48 ± 0.10	13.0 ± 1.0	0.2–2	− 0.47	− 0.47
RELACRE EAA[b]	Na	32.0 ± 3.0	30.7 ± 3.2	0.5–5	− 0.65	− 0.29
	K	7.70 ± 0.50	7.70 ± 0.86	0.4–2	0.00	0.00
	Mg	7.90 ± 0.50	8.30 ± 0.74	0.2–2	1.30	0.45
	Ca	36.0 ± 3.0	37.6 ± 2.9	0.2–2	0.80	0.38
	Cr	0.0200 ± 0.0030	0.0206 ± 0.0016	0.002–0.02	0.30	0.18
	Mn	0.0540 ± 0.0070	0.0510 ± 0.0033	0.004–0.02	− 0.75	− 0.39
	Fe	0.105 ± 0.020	0.103 ± 0.010	0.010–0.1	− 0.28	− 0.09
	Cu	0.096 ± 0.020	0.0960 ± 0.0097	0.004–0.02	0.00	0.00

[a] Aquacheck 1S, Round 485, May 2015—Soft water—Major Inorganic Components

[b] RELACRE EAA, 1st Round, June 2015, Drinking water

[c] Expanded uncertainties for a confidence level of 95 %

associated with the reference value, the detailed uncertainty models adequately described the quality of the measurements.

Conclusions

The developed methodology for estimating the mean recovery and respective standard uncertainty from the analysis of independent reference materials successfully pooled the recovery and uncertainty of various recovery tests into a single performance parameter. The estimated mean recovery and respective uncertainty allowed for the assessment of deviations from the ideal recovery relevant to ensure that the measurement results are traceable to the SI unit mg L^{-1}. The adequate identification of precision conditions affecting the estimated recovery allowed for the reliable estimation of the recovery uncertainty. For the analysis of samples with native analyte, before and after spiking, in the same run, measurement repeatability affects recovery estimation. For the analysis of certified reference materials or spiked samples with negligible native quantity, the intermediate precision should be considered for the estimation of the recovery uncertainty. The precision conditions affecting recovery estimation also influence the systematic effects assessed in the mean recovery. If the recovery estimation is affected by the measurement repeatability, the recovery reflects the combined effect of all systematic effects occurring in the measurement results. On the other hand, if the recovery estimation is affected by the intermediate precision, all the systematic effects, except the within-run systematic effect, are assessed by the mean recovery.

The developed methodology was successfully applied to the analysis of Na, K, Mg, Ca, Cr, Mn, Fe and Cu in waters of Aquacheck and RELACRE proficiency tests by ICP-OES.

Acknowledgements This work was supported by Fundação para a Ciência e Tecnologia (FCT) under project UID/QUI/00100/2013 and scholarship SFRH/BPD/110186/2015. The authors would like to acknowledge the useful and constructive suggestions of the anonymous reviewers.

References

1. Joint Committee for Guides in Metrology (2012) International vocabulary of metrology—basic and general concepts and associated terms (VIM). BIPM, Sèvres
2. Analytical Methods Committee (1995) Analyst 120:2303–2308
3. da Silva RJNB, Santos JR, Camões MFGFC (2006) Accred Qual Assur 10:664–671
4. Eurolab (2007) Measurement uncertainty revisited: alternative approaches to uncertainty evaluation. Eurolab, Paris
5. da Silva RJNB (2013) Water 5:1279–1302
6. da Silva RJNB, Williams A (Eds) (2015) Eurachem/CITAC guide: setting and using target uncertainty in chemical measurement. Eurachem, Europe
7. da Silva RJNB, Lino MJ, Santos JR, Camões MFGFC (2000) Analyst 125:1459–1464
8. da Silva RJNB, Figueiredo H, Santos JR, Camões MFGFC (2003) Anal Chim Acta 477:169–185
9. Correia AG, da Silva RJNB, Pedra F, Nunes MJ (2014) Accred Qual Assur 19:87–97
10. Miller J, Miller J (2005) Statistics and chemometrics for analytical chemistry, 5th edn. Pearson, London

11. Thompson M, Ellison SLR, Fajgelj A, Willetts P, Wood R (1998) Harmonised guidelines for the use of recovery information in analytical measurement. IUPAC, North Carolina
12. Magnusson B, Näykki T, Hovind H, Krysell M (2012) Handbook for calculation of measurement uncertainty in environmental laboratories, 3rd edn. Nordtest, Oslo
13. da Silva RJNB, Camões MFGFC (2010) Accred Qual Assur 15:691–704
14. Barwick VJ, Ellison SLR (1999) Analyst 124:981–990
15. Barwick VJ, Ellison SLR (2000) VAM project 3.2.1—part (d): protocol for uncertainty evaluation from validation data. LGC, London
16. Magnusson B, Örnemark U (2014) Eurachem guide: the fitness for purpose of analytical methods—a laboratory guide to method validation and related topics, 2nd edn. Eurachem, Europe
17. Rodrigues J, da Silva RJNB, Camões MFGFC, Oliveira CM (2015) Talanta 142:72–83
18. Brasil B, da Silva RJNB, Camões MFGFC, Salgueiro P (2013) Anal Chim Acta 804:287–295
19. Standards LGC (2015) Proficiency test report—aquacheck 1S, round 485, soft water—major inorganic components. LGC, London
20. RELACRE (2015) Relatório Final do Ensaio de Aptidão EAA Junho 2015. RELACRE, Lisboa
21. Commission Directive 2009/90/EC laying down, pursuant to Directive 2000/60/EC of the European Parliament and of the Council, technical specifications for chemical analysis and monitoring of water status. Off J Eur Union L201:36–38
22. Council Directive 98/83/EC on the quality of water intended for human consumption. Off J Eur Union L330:32–54
23. Agência Portuguesa do Ambiente (2016) Planos de Gestão de Região Hidrográfica 2016–2021. APA, Lisboa
24. da Silva RJNB (2016) Talanta 148:177–190
25. Analytical Methods Committee (1994) Analyst 119:2363–2366
26. da Silva RJNB, Santos JR, Camões MFGFC (2002) Analyst 127:957–963
27. ISO 5725-6:1994 Accuracy (trueness and precision) of measurement methods and results—part 6: use in practice of accuracy values. ISO, Geneva
28. Joint Committee for Guides in Metrology (2008) Evaluation of measurement data—guide to the expression of uncertainty in measurement. BIPM, Sèvres
29. Silva AMEV, da Silva RJNB, Camões MFGFC (2011) Anal Chim Acta 699:161–169
30. ISO 13529:2005 Statistical methods for use in proficiency testing by interlaboratory comparisons. ISO, Geneva