New approaches for interlaboratory comparisons analysis using dark uncertainty applied to radioactive materials

Marielle Crozet, Cedric Rivier, Valérie Lourenço, Séverine Demeyer

▶ **To cite this version:**

Marielle Crozet, Cedric Rivier, Valérie Lourenço, Séverine Demeyer. New approaches for interlaboratory comparisons analysis using dark uncertainty applied to radioactive materials. Talanta, 2022, 250, pp.123394. 10.1016/j.talanta.2022.123394 . cea-04155184

HAL Id: cea-04155184
https://cea.hal.science/cea-04155184v1

Submitted on 7 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

New approaches for interlaboratory comparisons analysis using dark uncertainty applied to radioactive materials

Marielle Crozet,[1] Cédric Rivier,[1] Valérie Lourenço,[2] Séverine Demeyer[3]

[1] CEA, DES, ISEC, DMRC, University of Montpellier, Marcoule, France
[2] Université Paris-Saclay, List, Laboratoire National Henri Becquerel (LNE-LNHB), 91120 Palaiseau, France
[3] Laboratoire national de métrologie et d'essais (LNE), Trappes, France

ABSTRACT

In order to further improve the management of contaminated materials in nuclear facilities subject to a decommissioning programme, as well as during post-accidental site remediation and clearance, the definition and selection of the most appropriate intervention scenarios producing well-characterized radioactive waste for which storage and disposal routes are clearly identified is needed.
As a step towards this accomplishment, we propose a methodology for the organization and analysis of coordinated interlaboratory comparisons (ILC) for the performance assessment and the uncertainty evaluation of available measurement techniques (methods and tools) of radioactive materials. This methodology is new for this type of comparison and demonstrated on the BR3 (Belgian Reactor 3, Belgian Nuclear Research Center, Mol) case study from the H2020 INSIDER project (2017-2021), for which barium 133, cobalt 60 and europium 152 are analysed with gamma spectroscopy in ILC, based either on irradiated concrete from the BR3 bioshield or from spiked concrete certified reference material (CRM).
On one hand, we show the advantage of organizing ILC on CRM for a more reliable uncertainty evaluation taking bias into account following ISO 21748:2017. But using CRM may be impossible due to their scarcity or too costly for performance assessment thus limiting the use of CRM in ILC in practice.
On the other hand, we show that for performance evaluation and monitoring, ILC can be alternately performed on reference materials provided that laboratories' uncertainties are reported and the most appropriate analysis of data is performed using dark uncertainty (excess variance) in the presence of inconsistent data.

Keywords: certified reference material (CRM) ; proficiency test (PT) ; Interlaboratory comparison (ILC) ; measurement uncertainty ; analysis of variance

1. Introduction

The nuclear industry is entering a period when its first installations approach the end of their lifetimes. The stakes are high: it is imperative to show that the life cycle of nuclear facilities can be brought to an end, leaving clean sites after use, and using optimized methods with limited costs. The dismantling and decommissioning (D&D) process of a site always follows the same steps: site characterization, elaboration of a scenario for the operations (including the preliminary decontamination of surfaces), cutting and dismantling operations, and management of waste and effluents. The European H2020 project INSIDER (Improved Nuclear SIte characterization for waste minimization in D&D operations under constrained EnviRonment) aims to develop a new integrated characterization methodology during D&D operations for nuclear power plants, post-accident land remediation, or nuclear facilities under constrained environments. Concerning the dismantling of reactors, the Belgian Nuclear Research Centre provided a case study: the characterization of the biological shield of Belgian Reactor 3 (BR3, Mol), which is made of irradiated heavy concrete.

The overall goal of the paper is to provide guidance on a smart organization and analysis of interlaboratory comparisons on radioactive materials to demonstrate the performance of measurement techniques and to evaluate the uncertainty associated with measurement results.

For these purposes, coordinated interlaboratory comparisons (ILCs) are performed on a certified reference material (CRM-ILC) concrete sample made from non-irradiated heavy concrete from the biological shield of BR3 as well as on two real irradiated concrete reference materials (RM-ILC) sampled from BR3 biological shield.

The organization of ILC on a CRM, as done in this study, is very valuable as it provides estimates of bias of laboratories that are used for a comprehensive uncertainty analysis including trueness and precision e.g. within ISO 21748:2017 [1] which we briefly recall. However, since it requires the production of the CRM (with the associated cost), CRM-ILC cannot frequently be organized in practice.

The organization of ILC on RMs is a fair compromise and should target a large variety of ILC. Since RM-ILC would yield an incomplete uncertainty evaluation w.r.t. CRM-ILC, we choose to focus in this paper on the use of

RM-ILC for proficiency assessment, where the consensus value is estimated from the participants and proficiency is assessed w.r.t. the consensus value. Due to inconsistency of data (the reported estimation of measurement uncertainties is not always fully mastered by the laboratories), dark uncertainty also called excess variance methods are required and guidance for the use of these methods is provided. To the best of our knowledge, these approaches are new for this type of measurements. In the paper, we choose to focus on the latest versions of derSimonian Laird algorithm and Bayesian analysis [2].

The paper is organized as follows. Section 2 presents the coordinated ILCs applied to radioactivity measurements and standard approaches for performance assessment and uncertainty evaluation. Section 3 provides guidance towards advanced statistical tools using dark uncertainty for the analysis of ILCs for performance assessment. Section 4 presents the results of all ILCs in terms of proficiency testing and uncertainty evaluation with a more general discussion on the scope of the results obtained from such interlaboratory studies. Conclusion is made in section 5.

2. Coordinated ILCs of radiological measurement methods

2.1. RM scarcity and use of ILCs in radioactivity measurements

The use of CRM is useful for the validation of analytical methods. However, radioactive CRMs are few in number, out of stock or close to it, and do not meet the specific needs of decommissioning, in terms of matrix and radioactive composition [3], [4], [5], [6]. Decommissioning requirements include the simultaneous measurement of several radionuclides and even their isotopic ratios. In this case, to validate the analytical methods, it may be necessary to use CRMs with a less than ideal composition (soil rather than steel or ion exchange resins [7]). Few CRM of matrices adapted for decommissioning needs, certified for their radioactivity contents, exist and even fewer whose values are traceable to the SI units [8].

The use of CRM from real materials taken from decommissioning sites is delicate because the samples are often inhomogeneous in nature. Typically, the steel of a reactor vessel or the concrete of its biological protection will have been the site of a neutron activation depending on the reactor flux, which varies according to the operating conditions of the reactor, to the distance to the fuel or to the elemental composition of the materials crossed, among others. The realization of CRMs from real samples will be adapted to the needs of decommissioning because their matrix will be representative as well as their radiological composition, but it will be all the more expensive because it may contain relatively low levels of radioactive elements to be measured (or types of emissions that are not very penetrating, such as pure alpha or beta emitters), and because it will require verification of the homogeneity of the final material.

To overcome these difficulties and to provide materials for calibration, inactive materials contaminated by radioactive spiking have been developed, but they are intended for verification of the criteria for free-release of final very-low-active waste (200-liter drum [9] or special Euro-pallet container [10]) and are therefore not suitable for on-site sampling measurements during decommissioning, intended to optimize the dismantling operations so as to minimize the wastes that are being produced.

To meet the needs of RMs in the field of D&D, it is tempting to make one from the samples taken on the construction sites and dedicated to destructive analyses in the laboratory [11], [12]. However, in the field of radiological analysis, often an ILC will allow verifying that the material is suitable for use as RM, from the consensus values obtained. However, this process can lead to high uncertainties, which will limit its use for the validation of analytical methods.

In volume, most of the wastes during D&D of a nuclear plant come from construction materials, especially concrete. In particular, it represents most of the low and intermediate level wastes. Accurate measurements of this matrix are thus needed in order to assign the material to the right waste treatment route, which depends on its activity levels. This is why the matrix chosen for the production of both RM and CRM was concrete. Moreover, the duration of the project being limited, non-destructive assays were favored because they are quicker and do not involve a prior dissolution step, which would increase measurement uncertainties. Gamma spectrometry is widely used, particularly in the early stages of decommissioning, because it allows relatively rapid identification and quantification of the radioactive content of a sample. It is adapted to the measurement of gamma-emitting radioactive elements with a sufficiently long half-life. All these aspects make it a method of choice for ILCs. In addition, as a non-destructive measurement method, gamma-ray spectrometry.

2.2. Purpose of coordinated ILCs

For CRM-ILC, laboratories are required to provide 5 individual measurements that will be processed for both performance assessment and uncertainty evaluation.

For RM-ILC, laboratories are required to provide a measurement result and its associated uncertainty for performance assessment.

Since the matrix, measurement technique and the activity level (for some radionuclides only) are the same for RM-ILC and CRM-ILC, the coordinated ILCs allow

- to compare the reported uncertainties from each laboratory with the global uncertainty estimated from all individual laboratory results. This information is very valuable for laboratories e.g. to have feedback on their uncertainty budgets ;
- to compare performance assessment on individual measurements (CRM-ILC) with performance assessment using reported uncertainties (on RM) to show that a more reliable performance assessment can be achieved by taking into account reported uncertainties (where the participants were asked to specify their coverage factor *k*).

2.3. Proficiency testing for performance assessment

Proficiency testing is the evaluation of participant performance against pre-established criteria by means of interlaboratory comparisons (ref 17043). The aim of PT is to compare a result on a proficiency test item with an assigned value, where a result is the average of all the measurement results x_j from a participant on the test item.

In order to assess proficiency in this study, two different situation types of proficiency test items were taken into account:

1) An ILC on a CRM for which the methods using performance scores according to ISO 13528:2015 [13] can be considered suitable: when using the concrete CRM as the test item, the assigned value x_{pt} for the proficiency test is the certified value and the standard uncertainty of the assigned value $u(x_{pt})$ is the uncertainty associated with the certified value. Standardized performance statistics (difference, z-score and ζ-score) are considered in this study. PT analyses were done using JMP SAS 14.00 statistical analysis software [14].

2) An ILC on real concretes where the assigned value is obtained from all the participants results (no CRM is used as a test item). This situation is a complex problem for which a variety of statistical approaches has been suggested [15] and is the object of section 3. In this paper, we present the DerSimonian-Laird (DL) procedure as well as the Bayesian procedure, and compare them with the uncertainty-weighted mean estimate in order to deal with excess variance. Performance is assessed with degrees of equivalence.

2.4. Accuracy of the measurement method

The accuracy of each of the methods used is characterized by its trueness and its precision. A method is true when there is no bias. To determine the trueness of a method, it is necessary to estimate its bias in relation to the certified reference value and test whether it is significant: in the present study, this is only possible for the test on the concrete CRM.

The precision of the method can be assessed by respecting the characteristic conditions described below:

– repeatability conditions; these are attained when the measurements are made by the same operator, on the same instrument, using a single method, within a short period of time in order to obtain measurements under conditions which are as identical as possible. In the present work, repeatability was taken to mean that the measurements were carried out by the same laboratory (identified with the respective same laboratory code).

– reproducibility conditions; these are attained when the following working conditions change: operator, instrument, slight modification in the method used, time (usually long time periods) between analyses, or any other cause which may add sources of variability.

2.4.1. Measurement uncertainty evaluation

According to ISO 21748:2017 [1], a general model for uncertainty evaluation can be expressed as

$$u^2(y) = s_R^2 + u^2\left(\hat{\delta}\right) + \sum c_i^2 u^2(x_i) \quad (1)$$

where s_R is the reproducibility standard deviation, $u(\hat{\delta})$ is the uncertainty associated with the bias of the method and $\sum c_i^2 u^2(x_i)$ is the sum of all of the effects due to other variations.

ISO 21748:2017 [1] standard extends the scope of ISO 5725-2:2019 [16] (using analysis of variance to estimate repeatability and reproducibility of the measurement method) when an estimate of trueness of the method is available, typically when a CRM is used as a test item.

For the CRM-ILC, we assume that there are no other steps to take into account during the analysis of an unknown sample (for example dissolution, additional dilution) with gamma spectrometry with a similar activity level, so that the third term of (1)

can be neglected and uncertainty may be estimated with the following equation:

$$\boldsymbol{u^2(y) = s_R^2 + u^2(\hat{\delta})} \quad (2)$$

2.4.2. Trueness

For a given analytical method, its trueness, δ, is the closeness of agreement between the best estimator of the result coming from a high number of results and a value considered to be the true value of the measurand, estimated by the certified reference value, which is equal to x_{pt}. δ is estimated by $\hat{\delta}$ and x_{ILC} is the best estimate coming from the laboratory results to the comparison.

$$\hat{\delta} = x_{ILC} - x_{pt} \quad (3)$$

Trueness evaluation requires the use of a CRM as the test item. The compatibility between x_{ILC} and the certified value x_{pt}, i.e. the absence of significative bias in the method, can be quantified by the normalized deviation E_n using (note that uncertainties at the denominator are not expanded as in ISO/IEC 17043:2010):

$$E_n = \frac{x_{ILC} - x_{pt}}{\sqrt{u^2(x_{ILC}) + u^2(x_{pt})}} = \frac{\hat{\delta}}{u(\hat{\delta})} \quad (4)$$

where $u(x_{ILC})$ is the standard uncertainty for the measurand estimation based on all the laboratory results, $u(x_{pt})$ is the standard uncertainty for the certified value, $\hat{\delta}$ is the estimate of the bias due to the method, and $u(\hat{\delta})$ is the estimate of the standard uncertainty associated with the bias of the method.

The normalized deviation values can be interpreted as follows:
$|E_n| \leqslant 2.0$: when the normalized deviation is between -2.0 and +2.0, there is no proven relevant difference between the two values x_{ILC} and x_{pt} : the estimated bias $\hat{\delta}$ is not significant. The risk associated with this conclusion is close to 5%.
$|E_n| \geqslant 2.0$: when the normalized deviation is less than -2.0 or greater than +2.0, there is a bias between the two values (and therefore in the method). The risk associated with this conclusion is close to 5%.

There are several ways to calculate the best measurand estimator based on laboratory results. For example, calculation of the mathematical mean, the robust mean, the weighted mean in general (weighted by a factor which is inversely proportional to the square of the uncertainty supplied by the laboratory), etc. In the present ILC on concrete CRM, the weighted mean was not used as some laboratories failed to adequately manage their uncertainty estimation. The normalized deviation was thus calculated by using the robust mean.

2.4.3. Precision

For a given analytical method, the individual measurement result x_{ij} is modeled according to:

$$x_{ij} = \mu + \alpha_j + \varepsilon_{ij},\ j = 1, \dots, k\ \ i = 1, \dots, n \text{ (and } x_j = \mu + \alpha_j + \frac{1}{n_j}\sum_{i=1}^{n_j} \varepsilon_{ij}) \quad (5)$$

where μ is the overall mean response, α_j is the effect of level j of laboratory factor and ε_{ij} is a random error term. This model is called a one-way random effects model, also frequently encountered for method validation as in [17].

The analysis of variance (anova) of this model is performed under the following hypotheses:
$\alpha_j \sim^{iid[1]} N(0, \sigma_L^2)$, $\varepsilon_{ij} \sim^{iid} N(0,\ \sigma_r{}^2)$ (homoscedasticity), and α_j, and ε_{ij} are pairwise independent. Significance testing of factors was performed with p-values obtained as the probability $P(F > F_{crit})$ where F is the value of a test statistic estimated on the data and F_{crit} is the value corresponding to a risk level $\alpha = 5\%$. A p-value less than

[1] iid : independent, identically distributed

$\alpha = 5\%$ indicates a significant effect. Under these assumptions (normality, homoscedasticity also known as homogeneity of variance and independence), variance components can be obtained from the anova sum of squares decomposition (for calculation details see [16]).

The reproducibility variance s_R^2 in (1) is the sum of the repeatability variance s_r^2 and the laboratory variance s_L^2

$$s_R^2 = s_r^2 + s_L^2 \quad (6)$$

where s_r^2 and s_L^2 are respectively the estimates of σ_r^2 and σ_L^2.

3. Analysis of RM-CIL for proficiency assessment using dark uncertainty

In the current study, specific statistical methods had to be used to deal with excess variance (also called heterogeneity or inconsistency of data) corresponding to situations where measured values are substantially more dispersed than what would be expected based on their reported uncertainties. Such methods produce results similar to those achieved by the uncertainty-weighted mean when there is no excess variance.

3.1. Uncertainty-weighted mean

When the $\{x_j\}$ are consistent among each other with respect to the quoted uncertainties $\{u_j\}$, the uncertainty-weighted mean, expressed as shown in (7) can be used to derive a combined estimate $\hat{\mu}_w$ of the measurement result and its associated uncertainty $u(\hat{\mu}_w)$.

$$\hat{\mu}_w = u^2(\hat{\mu}_w) \sum_{j=1}^{n} \frac{x_j}{u_j^2}, \; u(\hat{\mu}_w) = \left(\sum_{j=1}^{n} \frac{1}{u_j^2} \right)^{-1/2} \quad (7)$$

However, in order for this approach to be applicable, the results analysed need to be checked for consistency as the hypotheses underlying the use of the uncertainty-weighted mean do not hold true when working with inconsistent results. Inconsistency is usually verified by using the Cochran test of consistency. Under the consistency hypothesis, the following Q statistics follows a Chi-squared distribution with $n-1$ degrees of freedom [18]:

$$Q = \sum_{j=1}^{n} \frac{(x_j - \hat{\mu}_w)^2}{u_j^2} \sim Chi^2(n-1) \quad (8)$$

If the p-value, defined as $\Pr(Chi^2(n-1) > Q)$, is less than 0.05 then consistency is rejected at the level 5%. However, for small datasets the power of this test (capacity of the test to detect inconsistency) is usually poor and too sensitive for large datasets. For this reason, the rejection of consistency could be applied at a level of 10% in order to achieve a compromise [19].

It is worth noting that in "real world" situations, reported measurement results are often inconsistent, meaning that the uncertainty weighted mean cannot be considered applicable and random effects models need to be used instead.

3.2. Random effects model

The random effects model is written as:

$$x_j = \mu + \lambda_j + \varepsilon_j \quad (9)$$

where x_j is the measured value reported by laboratory j, μ is the overall arithmetic mean response, the $\{\lambda_j\}_{j=1,\ldots,n}$ are the laboratory effects (which are assumed to have a Gaussian distribution with a mean of 0 and a common standard deviation τ), the $\{\varepsilon_j\}_{j=1,\ldots,n}$ are random effects assumed Gaussian with mean 0 and standard deviation the reported standard uncertainties $\{u_j\}_{j=1,\ldots,n}$. The difference with model (5) lies in the specification of individually reported measurement uncertainties instead of a common residual variance parameter.

The parameter τ, often called dark uncertainty, accounts for heterogeneity amongst the measured values i.e. when the measured values are substantially more dispersed than would be expected from their stated laboratory-specific uncertainties [20]. As stated in [21] and [22], the rationale behind excess variance procedures is to enlarge reported variances by a common factor τ^2 which represents unexplained laboratory effects in order to avoid giving undue weight to results with a small reported uncertainty. In this paper, we focus on the DL

procedure and the Bayesian hierarchical procedure to estimate the parameters μ and τ of the random effects model **Erreur ! Source du renvoi introuvable.**). Proficiency is assessed in terms of degrees of equivalence, implemented as described in the NIST Consensus Builder (NICOB, version 1.4) [20]. For the benefit of the reader, all the necessary details regarding the statistical procedures utilized during this study, have been provided in this text. It is important to note that the classical DL procedure [23] has been enhanced to take into account the uncertainty on τ.

3.2.1. DerSimonian Laird procedure

The DL procedure [23] can be expressed as:

$$\hat{\mu}_{DL} = u^2(\hat{\mu}_{DL}) \sum_{j=1}^{n} \frac{x_j}{u_j^2 + \hat{\tau}_{DL}^2} \text{ with } u(\hat{\mu}_{DL}) = \left(\sum_{j=1}^{n} \frac{1}{u_j^2 + \hat{\tau}_{DL}^2} \right)^{-1/2} \qquad (10)$$

where $\hat{\tau}_{DL}^2 = \max\{0, \hat{\tau}_M^2\}$, $\hat{\tau}_M^2 = (Q - n + 1)/\left(\sum_{j=1}^{n} u_j^{-2} - \sum_{j=1}^{n} u_j^{-4} / \sum_{j=1}^{n} u_j^{-2}\right)$, and $Q = \sum_{j=1}^{n} u_j^{-2}\left(x_j - \hat{\mu}_w\right)$ is the Cochran statistics. When $\tau = 0$, the DL estimates are reduced to the uncertainty-weighted mean estimates. The Knapp-Hartung adjustment (KH) [22] can be used for building confidence intervals on these estimates, thus taking into account the unrecognized uncertainty around both the estimation of τ^2 and the reported u_j^2 by using the following pivotal quantity:

$$\frac{\hat{\mu}_{DL} - \mu}{\sqrt{u_{KH}^2(\hat{\mu}_{DL})}} \sim t_{n-1} \qquad (11)$$

where $u_{KH}^2(\hat{\mu}_{DL})$ is the KH estimate of the variance of $\hat{\mu}_{DL}$ defined as:

$$u_{KH}^2(\hat{\mu}_{DL}) = \left(\frac{\sum_{j=1}^{n} (x_j - \hat{\mu}_{DL})^2 / (u_j^2 + \hat{\tau}_{DL}^2)}{n-1} \right) u^2(\hat{\mu}_{DL}) \qquad (12)$$

The resulting confidence interval for μ at the level $1 - \alpha$ is $[\hat{\mu}_{DL} \pm t_{n-1,1-\frac{\alpha}{2}}\, u_{KH}(\hat{\mu}_{DL})]$ where $t_{n-1,1-\frac{\alpha}{2}}$ is the $1 - \alpha/2$ quantile of the t distribution with $n - 1$ degrees of freedom. A further enhancement of the method implemented in [20] (building on the KH adjustment) consists in sampling from the approximate distribution of the excess variance parameters using the parametric bootstrap Monte Carlo (PBMC) method as represented in Fig 1.

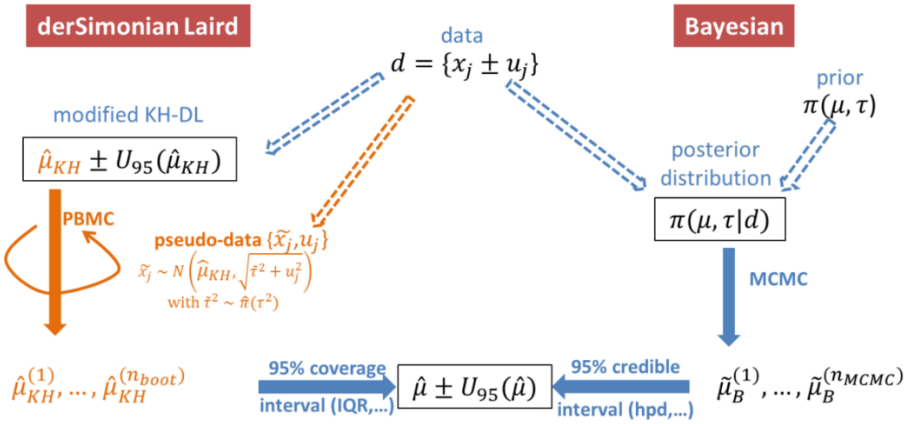


Fig. 1. Comparison of the DerSimonian-Laird and the Bayesian procedures to estimate consensus values.

3.2.2. Bayesian procedure

A Bayesian approach is used to estimate the posterior distributions of μ and τ given data $\mathrm{d} = \{\mathrm{x_j}, \mathrm{u_j^2}\}_{\mathrm{j=1,\ldots,n}}$. Bayesian inference combines one's prior knowledge about the parameters μ and τ with the information contained in the data. The result is the posterior distribution

$$\pi(\mu, \tau|d) = \frac{l(d|\mu, \tau)\pi(\mu, \tau)}{m(d)} \qquad (13)$$

where $l(d|\mu,\tau)$ is the likelihood of the data and $\pi(\mu,\tau)$ is the prior distribution expressing one's prior beliefs about μ,τ and $m(d)$ is a normalization constant. Equivalently, Bayes' formula can be expressed using the proportionality relation:

$$\pi(\mu,\tau|d) \propto l(d|\mu,\tau)\pi(\mu,\tau) \qquad (14)$$

A poorly informative prior can be chosen for μ, e.g. a Gaussian distribution with mean 0 and a very large standard deviation (say 10^5). Prior for variance components should be more carefully addressed in particular for low expected values. We follow here the recommendation [24] and assume that τ follows a half-Cauchy prior distribution parameterized by a scale parameter σ_τ as follows:

$$\pi(\tau|\sigma_\tau) = \frac{2}{\pi\sigma_\tau}\frac{1}{1+\tau^2/\sigma_\tau^2} \text{ if } \tau \geq 0 \text{ and } 0 \text{ otherwise} \qquad (15)$$

A recommendation of [20] is to take $\sigma_\tau = \mathrm{mad}(x_1,\ldots,x_n)$ where $\mathrm{mad}()$ is the median absolute deviation of the sample in the argument. Since the posterior distribution usually has no closed form, Markov Chain Monte Carlo (MCMC) methods [25] are employed to sample from the posterior distribution. These methods construct a sequence of dependent values which form a Markov chain with stationary distribution equal to the sought-after distribution. Amongst MCMC methods, the Metropolis-Hastings algorithm constitutes a popular class of methods as it only requires knowledge of the right hand part of equation (13) to sample from the posterior distribution. In this algorithm, the sequence of values is usually considered only after a first period of burn-in (e.g. discard the first 1000 simulations), and often the chains are thinned (e.g. only each 10^{th} value is used) in order to reduce the correlation between successive values. A general introduction to these methods can be found in [26], and [27], whilst [28] provides introductory example of their use in metrology.

3.2.3. Degrees of equivalence for performance assessment

Degrees of equivalence d_j and their 95% expanded uncertainties $U_{95}(d_j)$ are used to assess the agreement of laboratory values $x_j \pm U_{95}(x_j)$ with the consensus estimate $\hat{\mu}$ i.e. the performance of the method. In practice, degrees of equivalence are used to identify outliers with respect to the random effects model. The unilateral degree of equivalence (DoE) for laboratory j is defined as $d_j = x_j - \hat{\mu}$. As x_j is used to build the estimate $\hat{\mu}$, the covariance $\mathrm{cov}(x_j,\hat{\mu})$ between x_j and $\hat{\mu}$ should be estimated such that it has a reliable estimate of the uncertainty associated with each DoE according to the formula for the propagation of variances: $u^2(d_j) = u_j^2 + u^2(\hat{\mu}) - 2\mathrm{cov}(x_j,\hat{\mu})$. In order to avoid such complex computations, it was recommended in [29] that leave one out (LOO) estimates of the d_j defined as shown in equation (16), where $\widehat{\mu_{-j}}$ is the consensus estimate computed from all results but x_j, be considered.

$$d_j^{LOO} = x_j - \widehat{\mu_{-j}} \qquad (16)$$

In practice, LOO consists in repeating n times the estimation process. Once d_j^{LOO} is obtained either with the DL or the Bayesian procedure, measurement uncertainty $x_j \pm U_{95}(x_j)$ must still be taken into account. A schematic representation of the full procedure leading to the estimation of d_j and their associated uncertainty can be seen in Fig. 1. and Fig. 2Fig. .

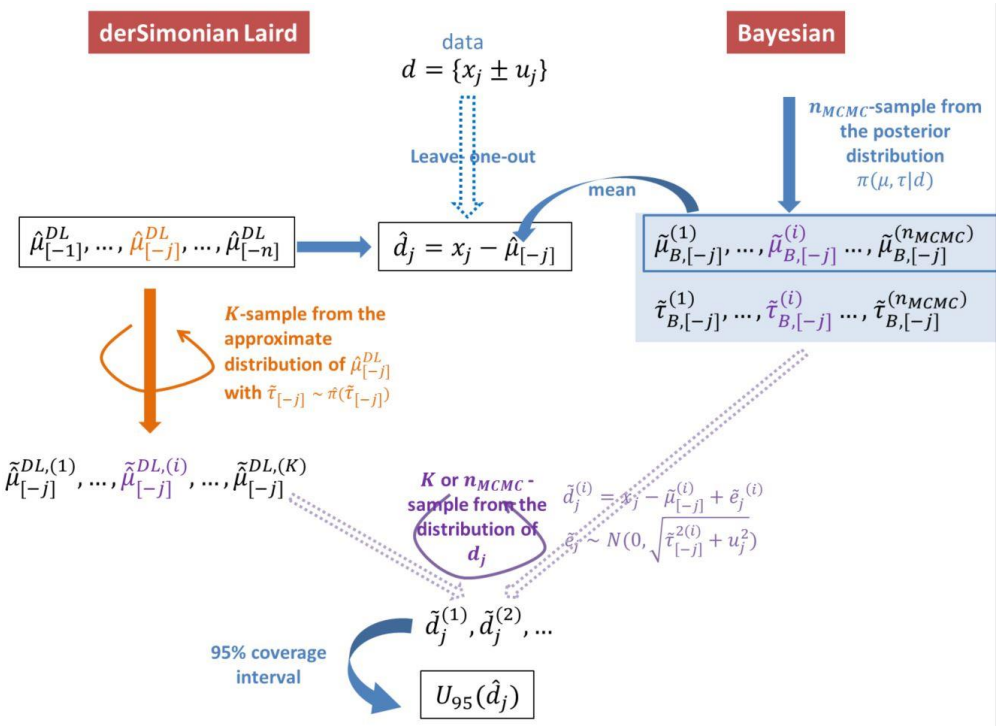


Fig. 2. Comparison of the DerSimonian-Laird and the Bayesian procedures to estimate unilateral degrees of equivalence

3.2.4. Comments on the DerSimonian-Laird and Bayesian procedures

In this paper, we present the enhanced version of the DL algorithm [23] as implemented in [20], with the KH adjustment and parametric bootstrap Monte Carlo (PBMC) method for the estimation of the consensus estimate, its associated uncertainty and coverage interval. These enhancements make the resulting uncertainty estimates comparable to the Bayesian estimates in their ability to take into account all the uncertainty sources, which is the usual justification for preferring Bayesian approaches. The resulting DL/KH/PBMC algorithm produces the same mean estimate of the consensus estimate as the initial DL algorithm from [23]. However, the enhanced algorithm differs significantly in terms of the uncertainty estimates in that the new, more conservative, estimates remain larger than those presented by the standard model. It is worth noting that, instead of producing analytical formulas for the mean and variance estimates, the new version relies on an iterative algorithm. For both the DL/KH/PBMC and Bayesian methods, the 95% coverage and credible intervals respectively are computed from simulated samples.

4. Results

In this section, we provide the analysis of the CRM-ILC and RM-ILC for proficiency testing with the most appropriate methods as explained in section 2.3. For RM-ILC we compare the excess variance consensus estimates with the uncertainty-weighted mean (called UW-mean), which is appropriate when data are consistent.
From CRM-ILC we are able to perform a complete uncertainty evaluation including bias. Individual reported standard uncertainties provided by laboratories for RM-ILC are compared with the global uncertainty computed from the CRM-ILC individual results (note that, as indicated in the text, the activity levels may differ).

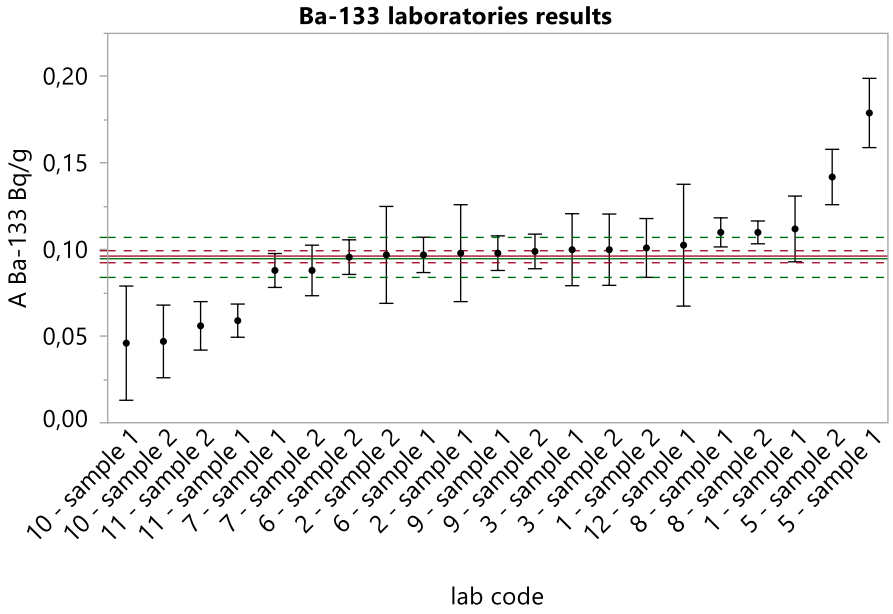
4.1. Analysis of CRM-ILC

4.1.1. Proficiency testing results

A summary of the PT results analysed according to section 2.3 can be found in Table 1.

For clarity, only the Ba-133 is presented in the main text of this article with the PT results for the other measurands - activity per unit mass (Bq/g) or mass activity in the following of Co-60, and of Eu-152 presented

in supplementary data (part A). Results for Ba-133 are displayed in Fig. 3 : $x_{pt} = x_{CRM} = 0.0960$ Bq/g ($u(x_{pt}) = u_{CRM} = 0.0018$ Bq/g ($k = 1$)). There is one outlier (sample 2 of code 12).



black dot: laboratory results ; bars: U_j (k=2); solid and dotted red line : x_{pt} and U_{pt} (k=2) ; solid and dotted green line: x^* and $U(x^*)$ (k=2)

Fig. 3. Ba-133 results for each of the two samples for PT analysis

For Ba-133 (at 0.0960 Bq/g), it can be seen that one result is an outlier. Additionally, of the remaining 23 results, only one z score is considered unacceptable (or action signal) whilst 8 zeta scores - corresponding to laboratory codes 5, 8, 10, and 11 each time for both of the studied samples - are unacceptable (or action signal). These results suggest that the laboratories are underestimating their uncertainty in the measurement of Ba-133 by gamma spectrometry.

For Co-60 (at 3.018 Bq/g), one result is an outlier. Additionally, of the remaining 23 results, only one z score is considered unacceptable whilst 2 zeta-scores - corresponding to laboratory codes 5 and 12 - are unacceptable. This suggests that these two laboratories are underestimating their uncertainty in the measurement of Co-60 by gamma spectrometry.

For Eu-152 (at 0.853 Bq/g), one result is an outlier. Additionally, all 23 remaining z scores are satisfactory whilst 4 zeta-scores - corresponding to laboratory codes 5 and 7 each time for both of the studied samples - are unacceptable. These results suggest that these two laboratories are underestimating their uncertainty in their measurement of Eu-152 by gamma spectrometry. Furthermore, it should be noted that laboratory 5 received an action signal for all of the gamma spectrometry measurements performed.

Table 1

Summary of the performance indicators for ILC on concrete CRM.

Lab code	Ba-133			Co-60			Eu-152		
	Dj%	zeta j	z j	Dj%	zeta j	z j	Dj%	zeta j	z j
1 - sample 1	17	1.7	0.7	7.4	0.9	1.5	-0.9	-0.1	-0.1
1 - sample 2	5.2	0.6	0.2	5.4	0.7	1.1	-4.0	-0.5	-0.3
2 - sample 1	2.1	0.1	0.1	5.7	0.6	1.2	7.9	0.6	0.6
2 - sample 2	1.0	0.1	0.0	1.7	0.2	0.3	0.8	0.1	0.1
3 - sample 1	4.2	0.4	0.2	4.4	0.5	0.9	6.1	0.6	0.5
3 - sample 2	4.2	0.4	0.2	6.4	0.7	1.3	10	1.0	0.8
4 - sample 1	_	_	_	_	_	_	_	_	_
4 - sample 2	_	_	_	_	_	_	_	_	_
5 - sample 1	86	8.2	3.8	16	3.4	3.3	27	5.9	2.2
5 - sample 2	48	5.6	2.1	7.7	1.7	1.6	17	3.9	1.4
6 - sample 1	1.0	0.2	0.0	2.8	0.5	0.6	-0.2	0.0	0.0
6 - sample 2	-0.3	-0.1	0.0	1.9	0.3	0.4	-2.8	-0.5	-0.2
7 - sample 1	-8.3	-1.5	-0.4	2.7	0.3	0.6	-23	-4.4	-1.8
7 - sample 2	-8.3	-1.1	-0.4	6.0	1.0	1.2	-18	-3.5	-1.5
8 - sample 1	15	3.1	0.6	0.1	0.0	0.0	2.0	0.2	0.2
8 - sample 2	15	3.7	0.6	-0.6	-0.2	-0.1	3.2	0.8	0.3
9 - sample 1	2.1	0.4	0.1	1.6	0.4	0.3	-1.4	-0.4	-0.1
9 - sample 2	3.1	0.6	0.1	3.0	0.8	0.6	-1.1	-0.3	-0.1
10 - sample 1	-52	-3.0	-2.3	-2.4	-0.4	-0.5	-18	-1.3	-1.5
10 - sample 2	-51	-4.6	-2.3	-5.0	-0.9	-1.0	-18.6	-1.7	-1.5
11 - sample 1	-39	-7.2	-1.7	-2.4	-0.4	-0.5	6.3	1.1	0.5
11 - sample 2	-42	-5.5	-1.8	-4.5	-0.7	-0.9	2.8	0.4	0.2
12 - sample 1	6.9	0.4	0.3	-8.5	-4.4	-1.7	-12	-2.9	-1.0
12 - sample 2	Outlier			Outlier			Outlier		

4.1.2. Accuracy of measurement methods on concrete CRM and uncertainty evaluation

Following the approach from section 2.4.2, the evaluation of the trueness of the gamma spectrometry method for each of the radionuclides of interest, based on the results of the comparison on the concrete CRM, is summarised in Table 2.

Table 2

Trueness evaluation of the gamma spectrometry method for Ba-133, Co-60 and Eu-152 measurement.

Radionuclide	Ba-133	Co-60	Eu-152
x_{pt}	0.0960	3.018	0.853
$u(x_{pt})$	0.0018	0.042	0.012
x_{ILC}	0.0954	3.084	0.844
$u(x_{ILC})$	0.0058	0.040	0.028
$\hat{\delta}$	-0.00059	0.066	-0.0094
$u(\hat{\delta})$	0.0061	0.058	0.031
E_n	**-0.1**	**1.1**	**-0.3**

The normalized deviation calculated for each of the radionuclides is less than 2 in absolute value, which indicates that the bias of the gamma spectrometry method is non-significant, which translates into $\hat{\delta} = 0$. As such, the method is not different from a true method for Ba-133, Co-60, and Eu-152 analysis in the ranges of this concrete CRM mass radioactivity.

Continuing the example of Ba-133, 4 of the 53 individual measurement results (corresponding to 3 results from code 10 and 1 result from code 5) can be considered outliers (according to the Grubbs test) and their distribution does not differ significantly from a normal distribution. It is therefore possible to carry out a variance analysis of these results (one way Anova, the studied factor is laboratory) whose outputs are presented in Fig. 4 and Table 3.

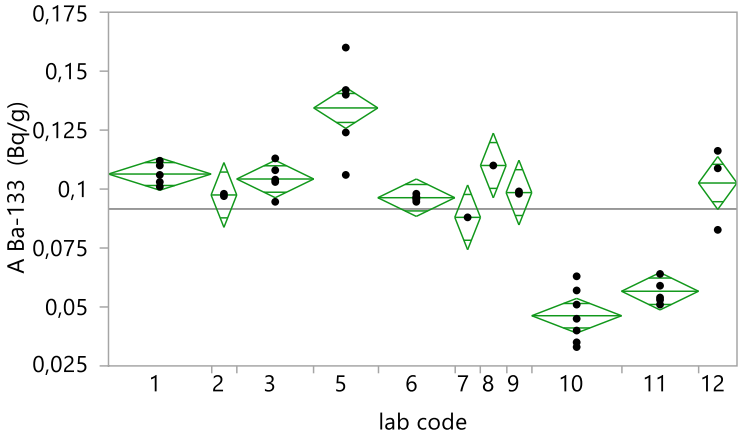


Fig. 4. Ba-133 by gamma spectrometry in CRM represented here as individual measurements from the two samples for variance analysis

Table 3

Ba-133 by gamma spectrometry - determination of the repeatability and the reproducibility uncertainties.

ANOVA **Ba-133 (Bq/g)**	total	between	within
	nk = 49	k = 11	
	global mean=	**0.0916**	
sum of the squares of the deviations	0.03842	0.03493	0.003498
number of degrees of freedom	48	10	38
variance	reprod. 0.0008744	lab 0.0007824	repeat. 0.00009200
uncertainty estimation (k = 1)	s_R	s_L	s_r
	0.030 **31%**	0.028 29%	**0.0096** **10%**

For measuring Ba-133 by gamma spectrometry, the relative intralaboratory standard uncertainty (repeatability) (10%) is lower than the relative standard uncertainty due to the laboratory factor (29%). From this, the reproducibility relative standard uncertainty for gamma spectrometry can be calculated to be 31%. The measurement uncertainty $u(y)$ associated with the measurement of Ba-133 by gamma spectrometry at a level around 0.096 Bq/g can be estimated as shown in Table 4, meaning a measurement standard uncertainty $u(y)$ equal to 0.030 Bq/g and a relative standard measurement uncertainty (k = 1) of 32% for a Ba-133 mass activity around 0.096 Bq/g.

Table 4

Ba-133 by gamma spectrometry - estimation of measurement uncertainty.

Ba-133 by gamma spec. **Bq/g**	u_{CRM}	$u(\hat{\delta})$	s_R	u(y) (k=1)
0.0960	0.0018	0.0061	0.0296	0.030 32%

Summary results of the other certified radionuclides can be found in Table 5.

The standard measurement uncertainty by gamma spectrometry is estimated at 5.2% for Co-60 at a level of about 3 Bq/g, at 12% for Eu-152 at a level of about 0.85 Bq/g, and at 32% for Ba-133 at a level of about 0.1 Bq/g.

It should be noted that the outliers eliminated before the analysis of variance come mainly from laboratory codes 5, 12, and - to a lesser extent - from laboratory codes 10 and 11; these codes are those that performed the worst with not satisfactory performance in the proficiency test.

Table 5

Characteristics of analytical methods on CRM and corresponding measurement uncertainty.

		analytical method characteristics				
			trueness	precision		
analytical method	Measurand : mass activity per unit mass	Value (Bq/g)	E_n	repeatability s_r	reproducibility s_R	measurement uncertainty u (k=1) (Bq/g)
gamma spectrometry	Ba-133	0.0960	**-0.1**	0.0096 (10%)	0.030 (31%)	0.030 (32%)
gamma spectrometry	Co-60	3.018	**1.1**	0.071 (2.3%)	0.14 (4.7%)	0.16 (5.2%)
gamma spectrometry	Eu-152	0.853	**-0.3**	0.035 (4.1%)	0.10 (12%)	0.10 (12%)

4.2. Analysis of RM-ILC for proficiency testing using excess variance procedures

4.2.1. RM low real concrete

Comparison between the DL procedure and the Bayesian procedure

A summary of the consensus estimates achieved using the DL procedure and the Bayesian procedure for the three measurands in the RM low real concrete sample is given in Table 6 and Table 7, respectively. It can be seen that there is a good agreement of the two methods. In order to allow for a comparison of the results, 100000 iterations of each procedure were performed for each radionuclide. Columns $\hat{\mu}$ and u($\hat{\mu}$) give the estimates of the assigned value and its associated uncertainty, respectively, with columns 2.5%($\hat{\mu}$) and 97.5%($\hat{\mu}$) respectively giving the lower and upper bounds of a 95% credible interval calculated from the samples. For the DL procedure, only a point estimate for τ in column $\hat{\tau}$ is available, whereas uncertainty and quantile estimates in columns u($\hat{\tau}$), 2.5%($\hat{\tau}$) and 97.5%($\hat{\tau}$) are obtained as by-products of the Bayesian procedure. The difference between the methods comes primarily from the excess variance estimation process.

In the remaining of the section, when DL and Bayesian results yield similar interpretation, only results obtained with DL are presented and only results for Ba-133 are displayed, for clarity. All the results obtained with the Bayesian analysis are displayed in the supplementary data part C. Results using DL for Co-60 and Eu-152 are displayed in Supplementary data part B.

Table 6

Results (Bq/g) for the consensus estimates using the DerSimonian-Laird procedure with 100 000 bootstrap simulations for RM low real concrete.

	$\hat{\mu}$	u($\hat{\mu}$)	2.5%($\hat{\mu}$)	97.5%($\hat{\mu}$)	$\hat{\tau}$
Ba-133	3.25	0.19	2.85	3.66	0.56
Co-60	0.0420	0.0010	0.0400	0.0450	0.0020
Eu-152	0.3159	0.0068	0.3023	0.3300	0.0124

Table 7

Results (Bq/g) for the consensus estimates using the Bayesian procedure with 100 000 MCMC simulations for RM low real concrete.

	$\hat{\mu}$	u($\hat{\mu}$)	2.5%($\hat{\mu}$)	97.5%($\hat{\mu}$)	$\hat{\tau}$	u($\hat{\tau}$)	2.5%($\hat{\tau}$)	97.5%($\hat{\tau}$)
Ba-133	3.25	0.23	2.79	3.72	0.68	0.18	0.42	1.12
Co-60	0.0424	0.0010	0.0406	0.0445	0.0018	$8.0\ 10^{-04}$	$8.0\ 10^{-04}$	0.0038
Eu-152	0.3155	0.0072	0.3012	0.3301	0.0126	0.0081	$9.0\ 10^{-04}$	0.0325

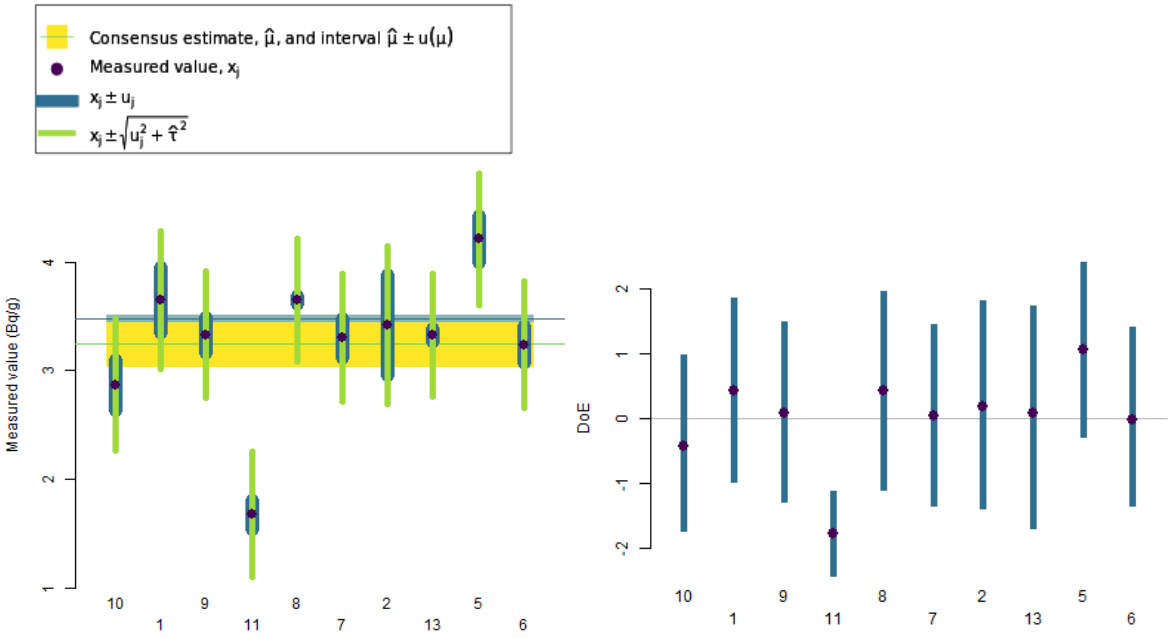
Interpretation of results for the three radionuclides

For Ba-133 (low) mass activity, the consensus graph in Fig. 5 (left) shows a huge shift of the UW-mean (blue line) from the DL consensus estimate (green line), that can be interpreted as the mean of laboratories 8 and 13 having the smallest uncertainties. In addition, the uncertainty associated with the UW-mean (blue band) is also too small, which makes UW-mean not reliable as a consensus. Taking into account dark uncertainty produces a more consensual estimate (green line) with a larger uncertainty band (in yellow) representative of the discrepancy of the results. The consensus plot also makes visible the "enlarged" uncertainties that are actually processed in the DL algorithm (green vertical bars) where the blue vertical bars represent the reported standard uncertainties. The result from laboratory 11 appears as an outlier, which is confirmed by the computation of its degree of equivalence in Fig. 5 (right), but we want to keep it in the analysis in case there is no instrumental/technical reason to doubt the reported result and uncertainty. So that excess variance approach can be considered as a robust method. For all other laboratories, the 95% credible intervals of the DOEs contain zero so that performance is achieved. In particular, for laboratories 5, 8 and 10 for which the zeta score was too large (a suspicion of underestimated uncertainties was raised section 4.1.1), correcting for too small reported uncertainties using dark uncertainty allows a more reliable performance evaluation.

For Co-60 (low) mass activity: the consensus graphs in supplementary data part B.a and part C.a show that the UW-mean is shifted towards the value of laboratory 8 which reports a very small uncertainty. Taking into account excess variance allows to retrieve a consensual value with an enlarged associated uncertainty with respect to the uncertainty associated with UW- mean, which is more representative of all of the reported results. Besides, the performance of laboratory 8 assessed with DOEs is nearly (but not) achieved whereas no suspicion was raised with the zeta score section 4.1.1, which suggests that the uncertainty at the level 0.04 Bg/g is underestimated but not the uncertainty reported at 3.018 Bq/g in the CRM. The results for laboratory 8 were nonetheless maintained in order to build the consensus value. Indeed, from the analysis of the consensus graph for Co-60, its effect on the consensus value is lowered with the excess variance approach, whereas the uncertainty-weighted estimate is almost confused with the result of this laboratory. On the contrary, laboratory 5 whose results were unacceptable at 3.018 Bq/g with both the z-score and the zeta score, now achieves performance at 0.04 Bg/g with the excess variance method which suggests that the uncertainty at the level 3.018 Bq/g was underestimated but not the uncertainty reported at 0.04 Bg/g.

For Eu-152 (low) mass activity: the consensus graph shows close results for the UW-mean and the excess-variance approach due to a lesser heterogeneity between results and reported uncertainties than those observed in

the results for Ba-133 and Co-60. Laboratories 5 and 7 whose results were unacceptable with the zeta score at 0.853 Bq/g (in the CRM) now achieve performance at 0.3155 Bq/g (in RM low), whereas laboratory 10 nearly reaches performance at 0.3155 Bq/g with the excess variance method, which suggests that the uncertainty of laboratory 10 at 0.3155 Bq/ may be slightly underestimated.



a) Consensus estimate b) Degrees of equivalence

Fig. 5. Left: Plot of the laboratory results and the consensus estimates using uncertainty-weighted mean and DL procedure. Right: Plot of the 95% credible interval for the degrees of equivalence for Ba-133 (low) mass activity.

4.2.2. RM high real concrete

Table 8 and Table 9 give a summary of both the consensus estimates and the dark uncertainty estimates achieved using the DL procedure and the Bayesian analysis respectively, for the three measurands in the RM high real concrete. As for the RM low real concrete, a good agreement of the two methods can be observed.

All the graphs referenced in this section are displayed in the supplementary data parts B and C for the DL and the Bayesian procedure respectively.

For Ba-133 high, the relative position of results is similar for all laboratories compared to results for Ba-133 low. The only difference with the low level analysis lies in the position of the UW-mean which is now at a central position but still with a too small associated uncertainty. Again, the effect of excess variance can be seen at is produces a more consensual estimate.For Co-60 high, the excess variance estimate is 0 with DL procedure, which is visible on the consensus graph in section B.c, where the vertical bars for the reported uncertainties and the vertical bars after taking dark uncertainty into account are the same. For this radionuclide, the difference between the DL and Bayesian procedure observed by comparing graphs in supplementary data part B.c. and part C.c. results from a significantly larger excess variance estimate obtained with the Bayesian approach. Since the Bayesian approach relies on the "true" distribution of the excess variance parameter given the model, Bayesian results should be preferred.

For Eu-152 high, a moderate effect of heterogeneity can be observed, which yields to a consensus uncertainty estimate not too far from UW-mean uncertainty.

Estimates and plots of degrees of equivalence are displayed in Supplementary material B.d. and C.d for the DL and the Bayesian procedure respectively.

In brief, the comparison with performance results on CRM yields to similar results and interpretations for Ba-133: only laboratory 11 does not achieve performance at 9.5316 Bq/g. All laboratories achieve performance for measuring Co-60 at 0.1272 Bq/g. For Eu-152, laboratory 10 does not achieve performance.

Table 8

Results for the consensus estimates using the DerSimonian-Laird procedure with 100 000 bootstrap simulations for RM high real concrete.

	$\hat{\mu}$	$u(\hat{\mu})$	2.5%($\hat{\mu}$)	97.5%($\hat{\mu}$)	$\hat{\tau}$
Ba-133	9.5316	0.7213	8.0932	10.9865	2.0873
Co-60	0.1272	0.0016	0.1239	0.1303	0
Eu-152	0.836	0.023	0.788	0.881	0.052

Table 9

Results for the consensus estimates using the DerSimonian-Laird procedure with 100 000 MCMC simulations for RM high real concrete.

	$\hat{\mu}$	$u(\hat{\mu})$	2.5%($\hat{\mu}$)	97.5%($\hat{\mu}$)	$\hat{\tau}$	u($\hat{\tau}$)	2.5%($\hat{\tau}$)	97.5%($\hat{\tau}$)
Ba-133	9.53	0.69	8.16	10.89	2.01	0.53	1.23	3.24
Co-60	0.1269	0.0017	0.1233	0.1301	0.0019	0.0016	1.0010^{-04}	0.0057
Eu-152	0.837	0.022	0.792	0.882	0.051	0.026	0.007	0.109

4.3. Discussion

In the concrete CRM, the gamma spectrometry method has been successfully assessed for its trueness for the measurement of Ba-133, Co-60, and Eu-152. As the matrix is the same as that of the real concretes, it can be stated that this applies also to the two real concretes studied. For Ba-133, the estimated reproducibility relative standard uncertainty decreases from about 31% at 0.1 Bq/g to about 6% at 10 Bq/g; for Co-60, from 9% at 0.05 Bq/g to 5% at 3 Bq/g; for Eu-152, the reproducibility relative standard uncertainty is about 10% at 0.3-0.8 Bq/g.

For Co-60 and Eu-152, the measurement uncertainties estimated from the results of the ILC for the CRM are consistent with the ones estimated by the laboratories:

- Co-60: relative standard uncertainty from ILC: 5% - relative standard uncertainty of the laboratories between 1% and 6%
- Eu-152: relative standard uncertainty from ILC: 12% - relative standard uncertainty of the laboratories between 3% and 17%

This tends to demonstrate that the evaluation of measurement uncertainty has been well controlled by the laboratories. This promising result can be attributed to the fact that the evaluation of measurement uncertainty in gamma spectrometry is well described in the literature [30], [31], [32].

As an example, considering that the measurement methods are the same for CRM and RM analysis, Fig.6 shows the comparison between reported uncertainties, excess variance uncertainties and uncertainties from the ILC on the CRM for the analysis of Eu-152 high for which the level of activity of RM high is close the level of activity of the CRM. From this example, it can be concluded that for all but one laboratory uncertainty from CRM is larger than the reported uncertainties which is to be expected.

However, for Ba-133, the conclusions are different. In that case, the measurement uncertainty estimated from the results of the ILC for the CRM (32%) is higher than most of the ones estimated by the laboratories (3 – 36%). This result can have several technical explanations:

- The very low activity of Ba-133 makes this radionuclide difficult to measure and makes its measurement very sensitive to background fluctuations. The background corrections applied by the laboratories can thus induce higher uncertainties than expected.
- The energies of Ba-133 gamma emission lines are lower than those of Co-60 and Eu-152 for which several high energy lines are available. The Ba-133 lines could therefore be more sensitive to matrix effects and efficiency variation. This is particularly relevant as the sample matrix (concrete) is different from that of the standard sources used for the calibration of the spectrometer.

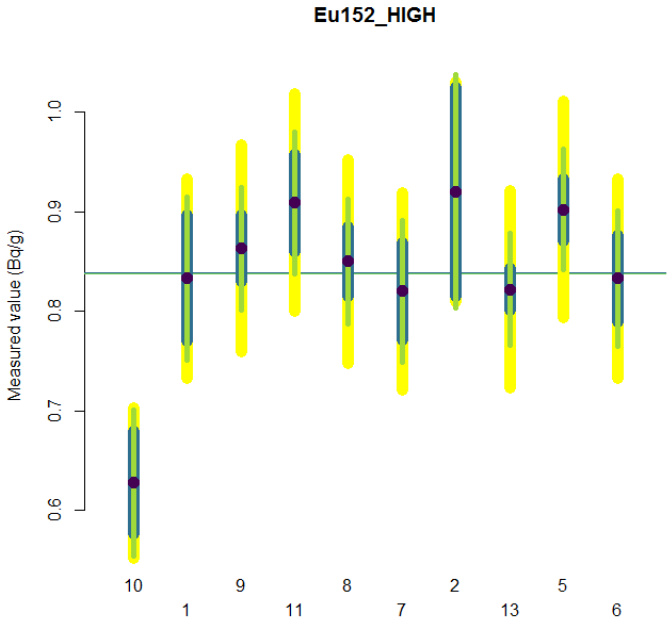


Fig. 6. Comparison between reported uncertainties (in blue), excess variance based uncertainties (in green) and uncertainties obtained from the comparison on the CRM (in yellow), all uncertainties are standard uncertainties. The horizontal blue line is the consensus estimate with the excess variance approach

5. Conclusion and perspectives

In this work, we propose a methodology for performance assessment and uncertainty evaluation using inter-laboratory comparisons. From a metrological perspective, the study confirms the importance of using a CRM to quantify bias and take into account the uncertainty due to bias in the uncertainty evaluation (otherwise restricted to the reproducibility variance). The resulting uncertainty can then be associated with the measured level of activity and could be used as the "default" uncertainty attributed to laboratories measuring the same activity with the same method on a similar matrix. For this kind of ILC, performance can be evaluated according to the ISO 13528:2015 [13] could be extended to assess the performance of laboratories reporting their own uncertainties in an ILC on reference (non certified) materials. For such ILCs, the phenomenon of excess variance is frequently encountered and must be taken into account (here with simulation based algorithms) for a sound performance assessment using degrees of equivalence and their associated 95% credible interval (instead of z-scores or zeta-scores for instance). The effect of considering excess variance (dark uncertainty) when supplied data is heterogeneous was studied, showing that when the data are close to homogeneous only a small difference (or even no difference, in case of full homogeneity) can be observed with the traditionally uncertainty-weighted mean. Finally, the study also showed that the comparison of the two approaches (ILCs on CRM and on RM with the same matrix) could be used to give insight into the completeness of the uncertainty budgets of participating laboratories. Indeed, the statistically based excess variance parameter (equivalent to 0 in the case of complete data homogeneity), which is usually used as an indicator of underestimated reported uncertainty, can be compared with the CRM based uncertainty. The dark uncertainty arising from excess variance approaches is an indicator of a measurement process not fully controlled. The sources of such uncertainty are always difficult to identify but should be investigated. In the case of gamma spectrometry, uncertainty sources such as background correction, calibration issues due to the differences between calibration sources and samples can cause such dark uncertainty. For the Ba-133 example, complementary work would be necessary to address the Ba-133 measurement issues. It would be worth studying and comparing the uncertainty estimation made by each laboratory in order to verify that the approaches were equivalent and that all sources of uncertainty were taken into account in the estimation. In a second step, it would be useful to estimate the impact of this benchmark measurement uncertainty estimated on the uncertainty associated with the radiological characterisation of the installation to be dismantled.

Acknowledgements

Fundind sources : The INSIDER project received funding from the Euratom Research and Training Programme 2014-2018 under grant agreement No 755554.

The authors thank the participants in ILCs : Astrid Barkleit (Helmholtz-Zentrum Dresden - Rossendorf e.V. (HZDR) Institute of Resource Ecology), Gianmarco Bilancia (Joint research Center (JRC) Ispra), René Brennetot (CEA/Direction des Energies, Institut des sciences Appliquées et de la Simulation pour les énergies bas carbone, Département de Physico-Chimie, Service d'Etudes Analytiques et de Réactivité des Surfaces, Laboratoire d'Analyse en soutien aux Exploitants), Sylvain Di Pasquale (Institute for Radioelements, Radioactivity measurement laboratory), Andrew Dobney (Belgian Nuclear Research Centre), Raquel Idoeta (University of the Basque Country, UPV/EHU, Laboratorio de Medidas de Baja Actividad, LMBA), Axel Klix (KIT/SUM, laboratory 28), Alexandra Nothstein (Karlsruher Institut für Technologie (KIT) Sicherheit und Umwelt (SUM), laboratory 26), Raf Van-Ammel (JRC Geel), Peter Volgyesi (Centre for Energy Research, Hungarian Academy of Sciences (MTA EK), Nuclear Security Department), Diana Walther (Dipl.-Chemiker, VKTA – Strahlenschutz, Analytik und Entsorgung Rossendorf e.V.).
The authors thank also the RMs producer i.e. Work Package 4 of INSIDER project lead by Ben Russel (National Physical Laboratory).

References

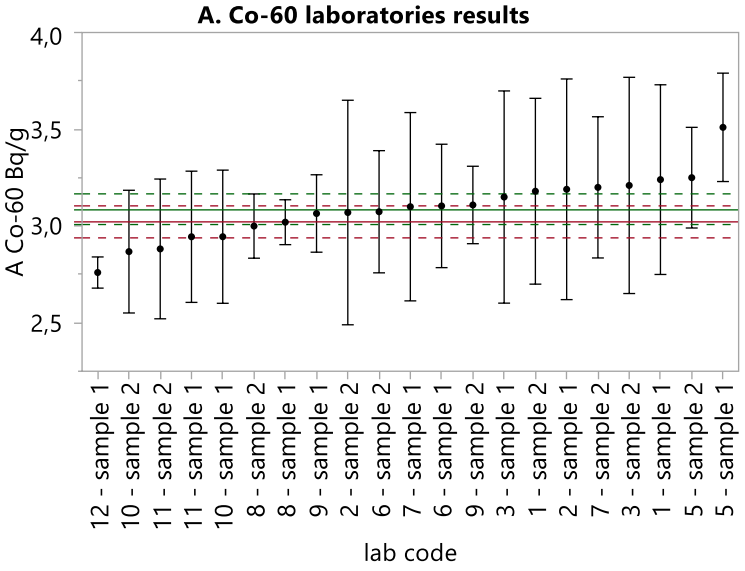
[1] ISO/TC69/SC6, ISO 21748:2017 Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation, (2017).
[2] NIST, NIST Consensus Builder, (2022). https://consensus.nist.gov/app/nicob.
[3] J.C.J. Dean, I. Adsley, P.H. Burgess, Traceability for measurements of radioactivity in waste materials arising from nuclear site decommissioning, Metrologia. 44 (2007) S140–S145. https://doi.org/10.1088/0026-1394/44/4/S18.
[4] S.J. Parry, Quality assurance in the nuclear sector, Radiochim. Acta. 100 (2012) 495–501. https://doi.org/10.1524/ract.2012.1958.
[5] K.G.W. Inn, C.M. Johnson, W. Oldham, S. Jerome, L. Tandon, T. Schaaff, R. Jones, D. Mackney, P. MacKill, B. Palmer, D. Smith, S. LaMont, J. Griggs, The urgent requirement for new radioanalytical certified reference materials for nuclear safeguards, forensics, and consequence management, J. Radioanal. Nucl. Chem. 296 (2013) 5–22. https://doi.org/10.1007/s10967-012-1972-y.
[6] E. Braysher, B. Russell, S.M. Collins, E.M. van Es, R. Shearman, F.D. Molin, D. Read, M. Anagnostakis, R. Arndt, A. Bednár, T. Bituh, J.P. Bolivar, J. Cobb, N. Dehbi, S. Di Pasquale, C. Gascó, C. Gilligan, P. Jovanovič, A. Lawton, A.M.J. Lees, A. Lencsés, L. Mitchell, I. Mitsios, B. Petrinec, J. Rawcliffe, M. Shyti, J.A. Suárez-Navarro, S. Suursoo, E. Tóth-Bodrogi, T. Vaasma, L. Verheyen, J. Westmoreland, G. de With, Development of a reference material for analysing naturally occurring radioactive material from the steel industry, Anal. Chim. Acta. 1141 (2021) 221–229. https://doi.org/10.1016/j.aca.2020.10.053.
[7] L. Zhu, X. Hou, J. Qiao, Determination of 135Cs concentration and 135Cs/137Cs ratio in waste samples from nuclear decommissioning by chemical separation and ICP-MS/MS, Talanta. 221 (2021) 121637. https://doi.org/10.1016/j.talanta.2020.121637.
[8] T. Altzitzoglou, A. Bohnstedt, Characterisation of the IAEA-375 Soil Reference Material for radioactivity, Appl. Radiat. Isot. 109 (2016) 118–121. https://doi.org/10.1016/j.apradiso.2015.11.053.
[9] J. Dean, A UK comparison for measurements of low levels of gamma-emitters in waste drums, Appl. Radiat. Isot. 67 (2009) 678–682. https://doi.org/10.1016/j.apradiso.2009.01.009.
[10] J. Suran, P. Kovar, J. Smoldasova, J. Solc, L. Skala, D. Arnold, S. Jerome, P. de Felice, B. Pedersen, T. Bogucarska, F. Tzika, R. van Ammel, New high-throughput measurement systems for radioactive wastes segregation and free release, Appl. Radiat. Isot. 130 (2017) 252–259. https://doi.org/10.1016/j.apradiso.2017.09.043.
[11] A. Harms, C. Gilligan, Development of a neutron-activated concrete powder reference material, Appl. Radiat. Isot. 68 (2010) 1471–1476. https://doi.org/10.1016/j.apradiso.2009.11.031.
[12] A. Leskinen, C. Gautier, A. Räty, T. Kekki, E. Laporte, M. Giuliani, J. Bubendorff, J. Laurila, K. Kurhela, P. Fichet, S. Salminen-Paatero, Intercomparison exercise on difficult to measure radionuclides in activated concrete—statistical analysis and comparison with activation calculations, J. Radioanal. Nucl. Chem. 329 (2021) 945–958. https://doi.org/10.1007/s10967-021-07824-7.
[13] ISO/TC69/SC6, ISO 13528:2015 Statistical methods for use in proficiency testing by interlaboratory comparison, (2015).
[14] JMP® 14.0.0 (Statistical Discovery) commercial software from SAS Institute Inc., (2022).

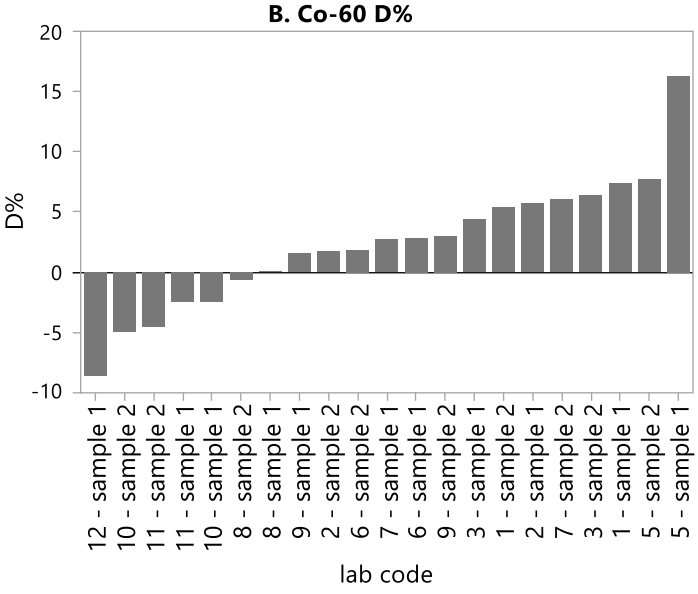
[15] CCQM, CCQM Guidance note: Estimation of a concensus KCRV and associated degrees of equivalence, (2013). http://www.bipm.org/cc/CCQM/Allowed/19/CCQM13-22_Consensus_KCRV_v10.pdf.
[16] ISO/TC69/SC6, ISO 5725-2:2019 Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method, (2019).
[17] T. Saffaj, B. Ihssane, A Bayesian approach for application to method validation and measurement uncertainty, Talanta. 92 (2012) 15–25. https://doi.org/10.1016/j.talanta.2011.11.077.
[18] M.G. Cox, The evaluation of key comparison data, Metrologia. 39 (2002) 589.
[19] J.P. T, Higgins, Vivian, T. James, C. Jacqueline, C. Miranda, L. Tianjing, P. Matthew, Welch, Cochrane Handbook for Systematic Reviews of Interventions, version 6., Wiley, Chichester UK, 2021. https://doi.org/10.1002/9781119536604.
[20] A.A. Koepke, T. Lafarge, A. Possolo, B. Toman, Consensus Builder User ' s Manual., (2020).
[21] C. Rivier, M. Désenfant, M. Crozet, C. Rigaux, D. Roudil, B. Tufféry, A. Ruas, Use of an excess variance approach for the certification of reference materials by interlaboratory comparison, Accredit. Qual. Assur. 19 (2014) 269–274. https://doi.org/10.1007/s00769-014-1066-3.
[22] R.C.M. Aert, D. Jackson, A new justification of the Hartung-Knapp method for random-effects meta-analysis based on weighted least squares regression, Res. Synth. Methods. 10 (2019) 515–527. https://doi.org/10.1002/jrsm.1356.
[23] R. DerSimonian, N. Laird, Meta-analysis in clinical trials, Control. Clin. Trials. 7 (1986) 177–188. https://doi.org/10.1016/0197-2456(86)90046-2.
[24] A. Gelman, Prior distributions for variance parameters in hierarchical modelsNo Title, Bayesian Anal. 1 (2006) 215–533.
[25] V. Roy, Convergence Diagnostics for Markov Chain Monte Carlo, Annu. Rev. Stat. Its Appl. 7 (2020) 387–412. https://doi.org/10.1146/annurev-statistics-031219-041300.
[26] C.P. Robert, G. Casella, Monte Carlo Statistical Methods, Springer New York, New York, NY, 2004. https://doi.org/10.1007/978-1-4757-4145-2.
[27] J.M. Bernardo, A.F.M. Smith, eds., Bayesian Theory, John Wiley & Sons, Inc., Hoboken, NJ, USA, 1994. https://doi.org/10.1002/9780470316870.
[28] K. Klauenberg, C. Elster, Markov chain Monte Carlo methods: an introductory example, Metrologia. 53 (2016) S32–S39. https://doi.org/10.1088/0026-1394/53/1/S32.
[29] D.L. Duewer, K.W. Pratt, C. Cherdchu, N. Tangpaisarnkul, A. Hioki, M. Ohata, P. Spitzer, M. Máriássy, L. Vyskočil, "Degrees of equivalence" for chemical measurement capabilities: primary pH, Accredit. Qual. Assur. 19 (2014) 329–342. https://doi.org/10.1007/s00769-014-1076-1.
[30] M.C. Lépy, A. Pearce, O. Sima, Corrigendum: Uncertainties in gamma-ray spectrometry (2015 Metrologia 52 S123–45), Metrologia. 54 (2017) 883–883. https://doi.org/10.1088/1681-7575/aa853b.
[31] ISO/TC85/SC2, ISO 18589-3:2018 Measurement of radioactivity in the environment — Soil — Part 3: Test method of gamma-emitting radionuclides using gamma-ray spectrometry, (2018).
[32] ISO/TC147/SC3, ISO 10703:2021 Water quality — Gamma-ray emitting radionuclides — Test method using high resolution gamma-ray spectrometry, (2021).

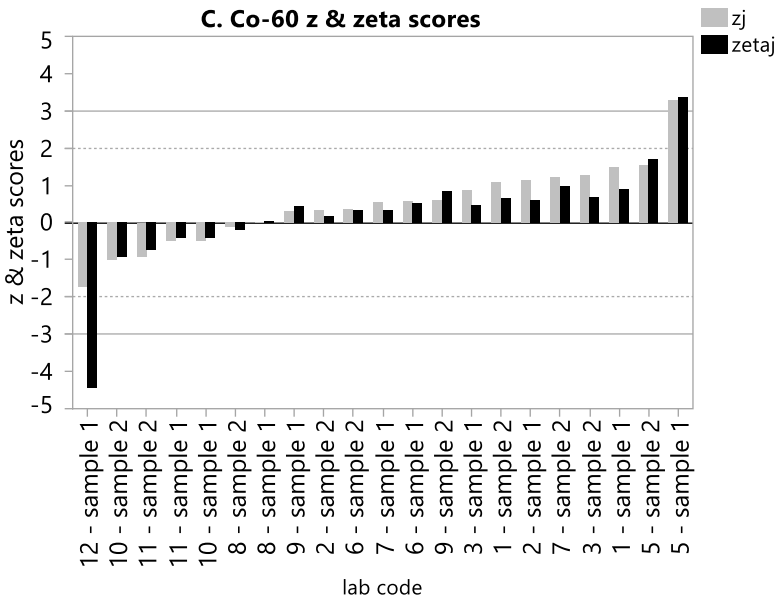
Supplementary data

A. Proficiency test on concrete CRM - measurands: mass activity of Co-60 and of Eu-152

Co-60 mass activity: $x_{pt} = x_{CRM} = 3.018$ Bq/g ($u(x_{pt}) = u_{CRM} = 0.042$ Bq/g $(k = 1)$)
1 outlier



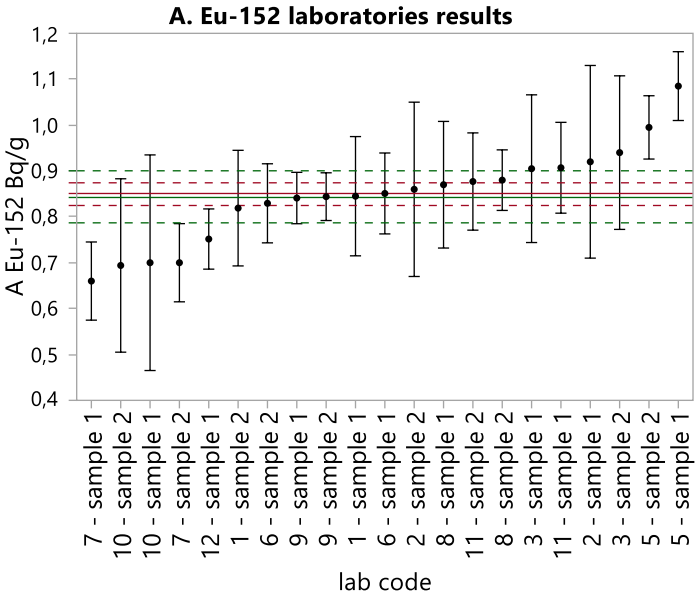


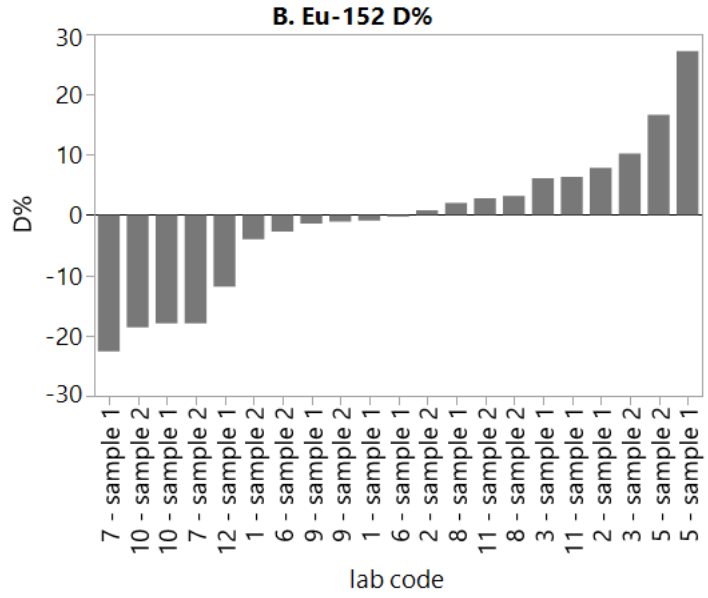


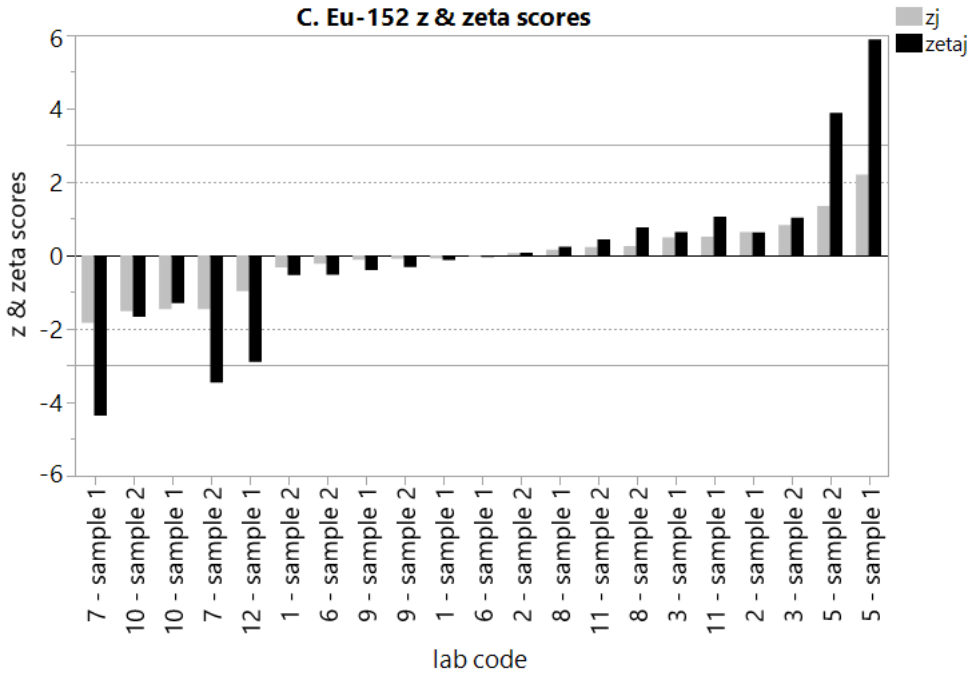
A: black dot: laboratory results ; bars: U_j (k=2); solid and dotted red line : x_{pt} and U_{pt} (k=2) ; solid and dotted green line: x^* and $U(x^*)$ (k=2)

Fig. 7. Laboratory results (A), deviation (B) and z and zeta scores (C) for Co-60 mass activity in concrete CRM.

Eu-152 mass activity: $x_{pt} = x_{CRM} = 0.853$ Bq/g ($u(x_{pt}) = u_{CRM} = 0.012$ Bq/g ($k = 1$))
1 outlier







A: black dot: laboratory results ; bars: U_j (k=2); solid and dotted red line : x_{pt} and U_{pt} (k=2) ; solid and dotted green line: x^* and $U(x^*)$ (k=2)

Fig. 8. Laboratory results (A), deviation (B) and z and zeta scores (C) for Eu-152 mass activity in concrete CRM.

B. Results with DerSimonian Laird procedure

a. DerSimonian Laird estimates for consensus estimates for low level real concrete

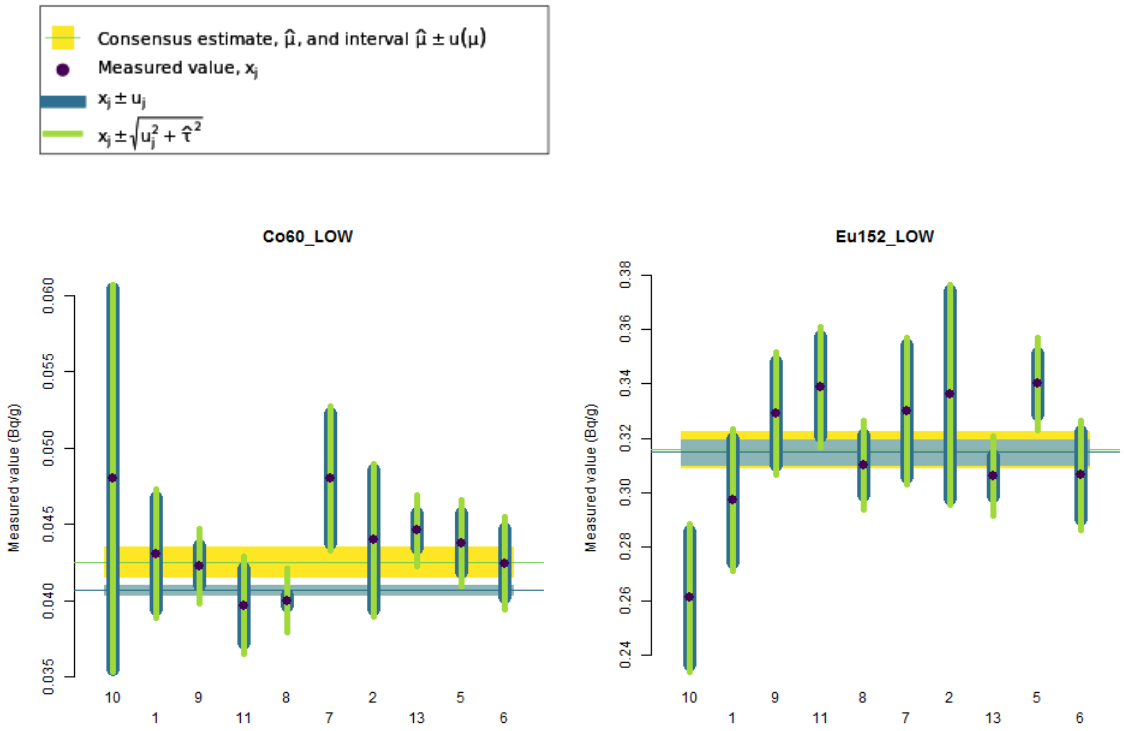


Fig. 9. Low level real concrete: DerSimonian-Laird estimates for Co-60 and Eu-152 mass activity.

b. DerSimonian-Laird estimates for degrees of equivalence for low level

Lab	DoE.x	DoE.U95	DoE.Lwr	DoE.Upr
10	0.0055	0.0251	-0.0199	0.0303
1	$6.00\ 10^{-04}$	0.0086	-0.008	0.0092
9	$-3.00\ 10^{-04}$	0.0059	-0.0062	0.0057
11	-0.0032	0.0068	-0.01	0.0036
8	-0.0034	0.0018	-0.0052	-0.0015
7	0.0057	0.0095	-0.0038	0.0151
2	0.0015	0.0101	-0.0085	0.0116
13	0.0029	0.004	-0.0011	0.0068
5	0.0014	0.0061	-0.0047	0.0075
6	$-1.00\ 10^{-04}$	0.0067	-0.0068	0.0065

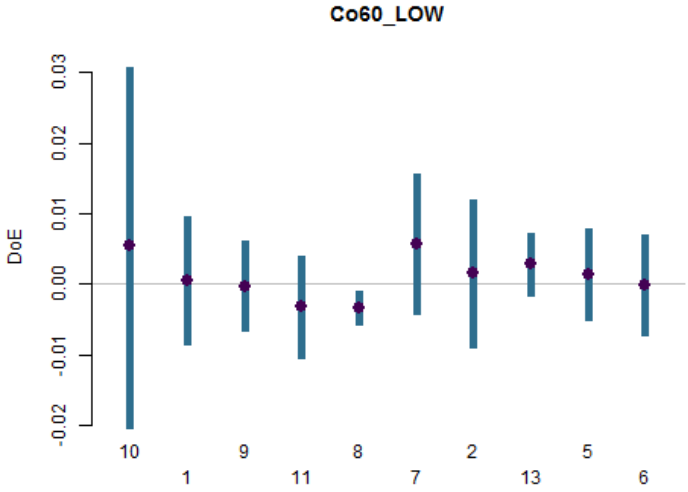


Fig. 10. Low level real concrete: DerSimonian-Laird estimate of degrees of equivalence for Co-60.

Lab	DoE.x	DoE.U95	DoE.Lwr	DoE.Upr
10	-0.0571	0.0533	-0.1109	-0.0044
1	-0.02	0.0544	-0.075	0.0336
9	0.0145	0.0488	-0.0343	0.0633
11	0.0255	0.0459	-0.0215	0.0703
8	-0.0067	0.0421	-0.0483	0.0357
7	0.0152	0.0574	-0.043	0.0719
2	0.0209	0.0816	-0.0606	0.1025
13	-0.012	0.0375	-0.0492	0.026
5	0.0294	0.0314	-0.0026	0.0605
6	-0.0105	0.0448	-0.0548	0.0346

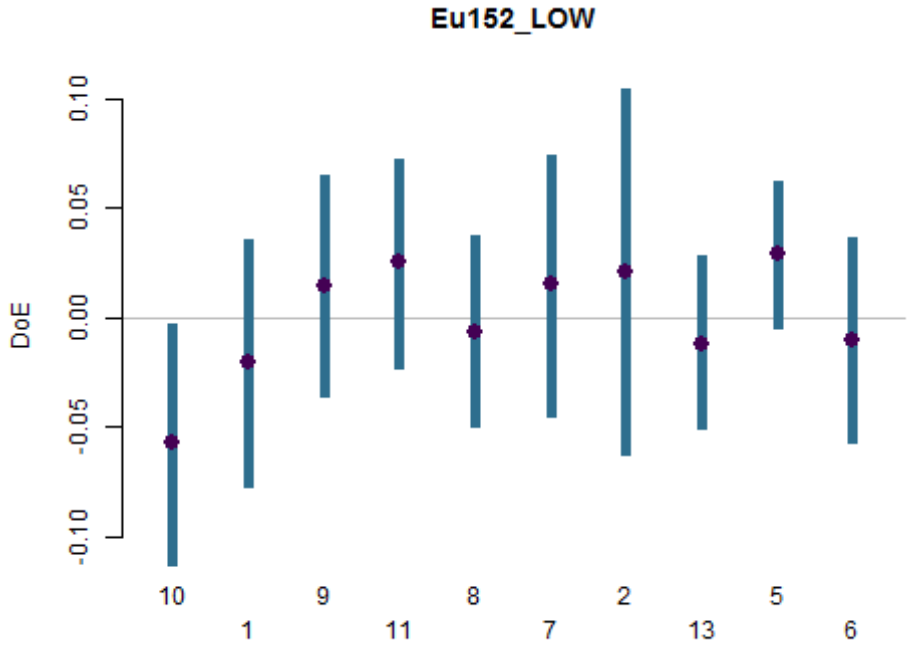


Fig. 11. Low level real concrete: DerSimonian-Laird estimate of degrees of equivalence for Eu-152.

c. DerSimonian-Laird estimates for consensus estimates for high level

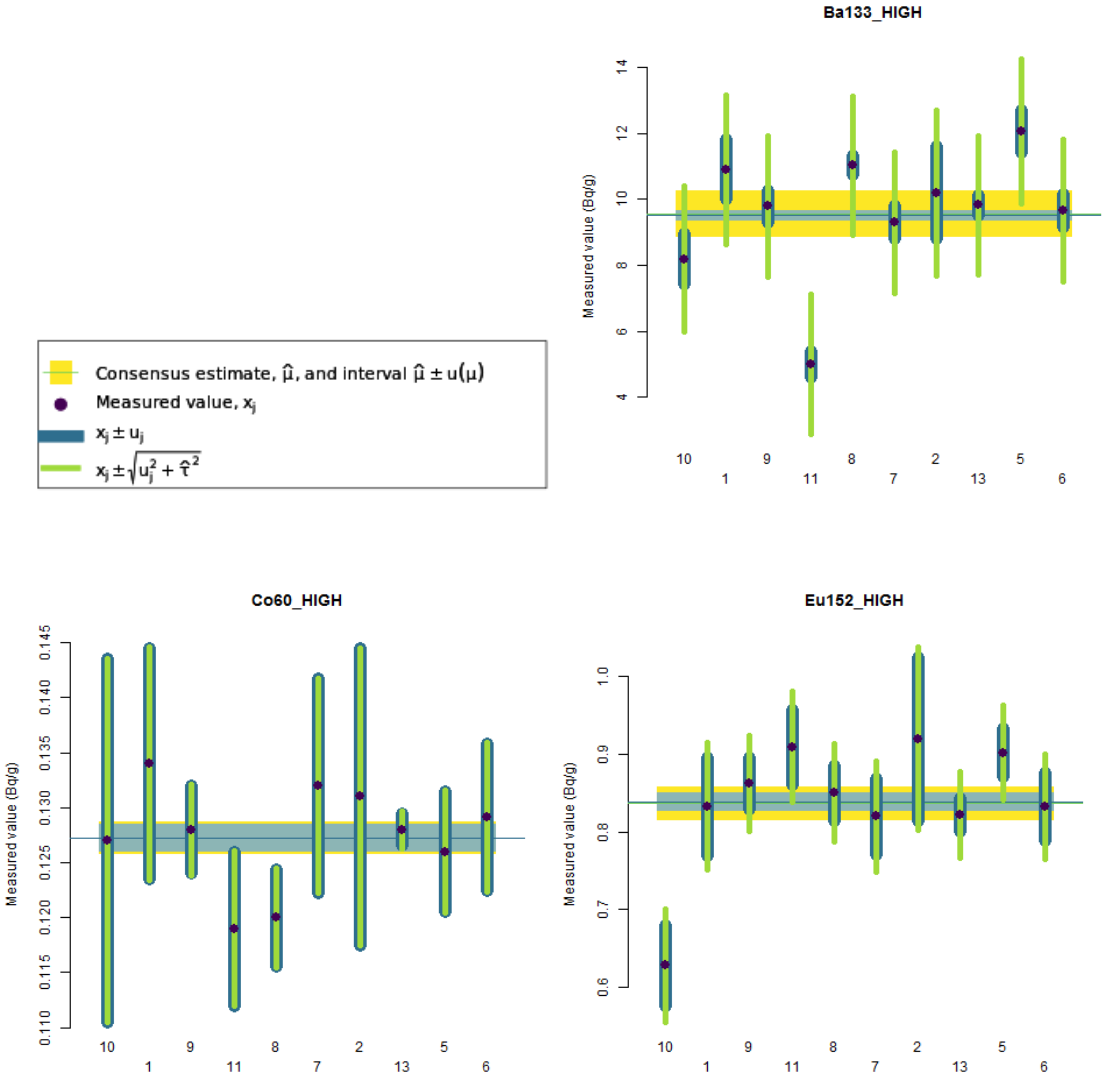


Fig. 12. High level real concrete: DerSimonian-Laird estimates for Ba-133, Co-60 and Eu-152 mass activity.

d. DerSimonian estimates for degrees of equivalence for high level

Lab	DoE.x	DoE.U95	DoE.Lwr	DoE.Upr
10	-1.5193	4.7638	-6.2196	3.2933
1	1.4755	4.8786	-3.5101	6.249
9	0.2534	4.8064	-4.542	5.0832
11	-5.1098	1.9962	-7.0942	-3.103
8	1.6438	4.7067	-3.0659	6.3259
7	-0.2943	4.8387	-5.123	4.5531
2	0.6914	5.1582	-4.4878	5.8322
13	0.2821	5.2281	-4.9041	5.54
5	2.7805	4.5375	-1.7142	7.3522
6	0.1018	4.8278	-4.7095	4.9521

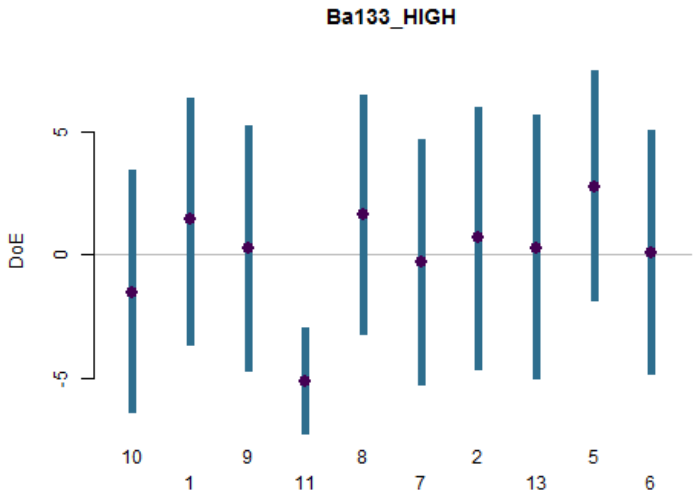


Fig. 13. High level real concrete: DerSimonian-Laird estimate of degrees of equivalence for Ba-133.

Lab	DoE.x	DoE.U95	DoE.Lwr	DoE.Upr
10	$-3.00\ 10^{-04}$	0.0326	-0.032	0.033
1	0.0068	0.0212	-0.014	0.0282
9	$8.00\ 10^{-04}$	0.0085	-0.0077	0.0094
11	-0.0085	0.014	-0.0224	0.0058
8	-0.0078	0.0093	-0.0171	0.0015
7	0.0048	0.0193	-0.0144	0.0242
2	0.0038	0.0268	-0.0226	0.031
13	0.0024	0.0054	-0.003	0.0078
5	-0.0013	0.0114	-0.0128	0.0098
6	0.0019	0.0135	-0.0116	0.0154

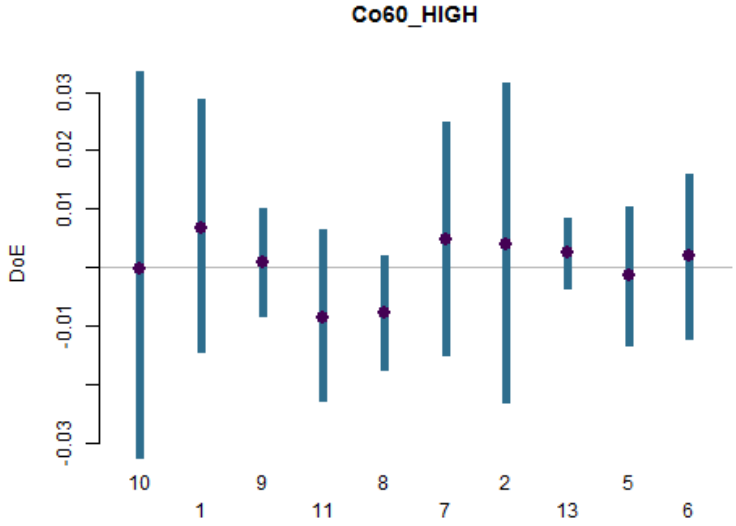


Fig. 14. High level real concrete: DerSimonian-Laird estimate of degrees of equivalence for Co-60.

Lab	DoE.x	DoE.U95	DoE.Lwr	DoE.Upr
10	-0.2227	0.1078	-0.3298	-0.1145
1	-0.0039	0.1719	-0.1792	0.1655
9	0.0301	0.144	-0.116	0.1733
11	0.0796	0.1472	-0.0686	0.2257
8	0.0153	0.1483	-0.1375	0.1607
7	-0.0183	0.1551	-0.1752	0.135
2	0.0863	0.2356	-0.1506	0.3201
13	-0.0167	0.1464	-0.1642	0.1289
5	0.0748	0.1299	-0.056	0.2033
6	-0.0041	0.1509	-0.1538	0.1479

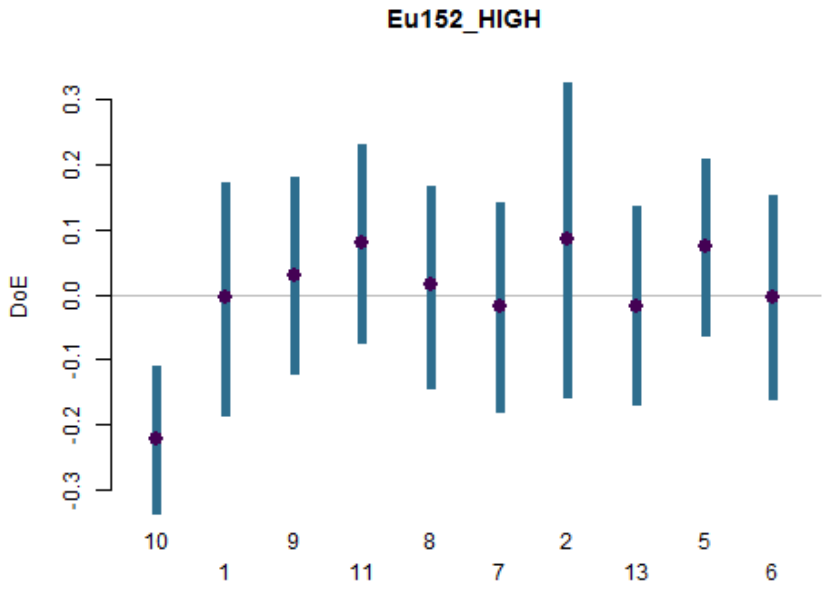


Fig. 15. High level real concrete: DerSimonian-Laird estimate of degrees of equivalence for Eu-152.

C. Results with the Bayesian procedure

a. Bayesian estimates for consensus estimates for low level

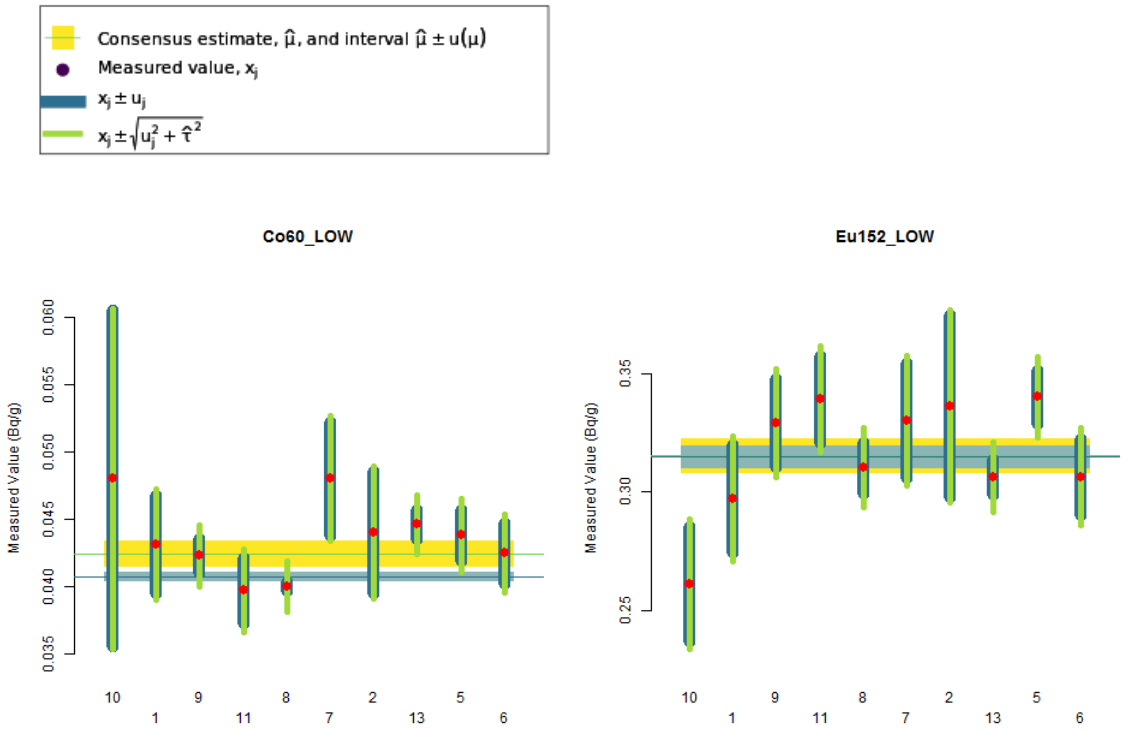


Fig. 16. Low level real concrete: Bayesian estimates for Co-60 and Eu-152 mass activity.

b. Bayesian estimates for degrees of equivalence for low level

Lab	DoE.x	DoE.U95	DoE.Lwr	DoE.Upr
10	0.0056	0.0244	-0.0197	0.0295
1	$7.00\ 10^{-04}$	0.0087	-0.0083	0.0091
9	$-2.00\ 10^{-04}$	0.0055	-0.0058	0.0052
11	-0.003	0.0068	-0.0099	0.0037
8	-0.0033	0.0032	-0.0066	$-2.00\ 10^{-04}$
7	0.0058	0.0093	-0.0032	0.0155
2	0.0017	0.01	-0.0085	0.0114
13	0.003	0.0046	-0.002	0.0071
5	0.0015	0.006	-0.0047	0.0074
6	0	0.0063	-0.0062	0.0063

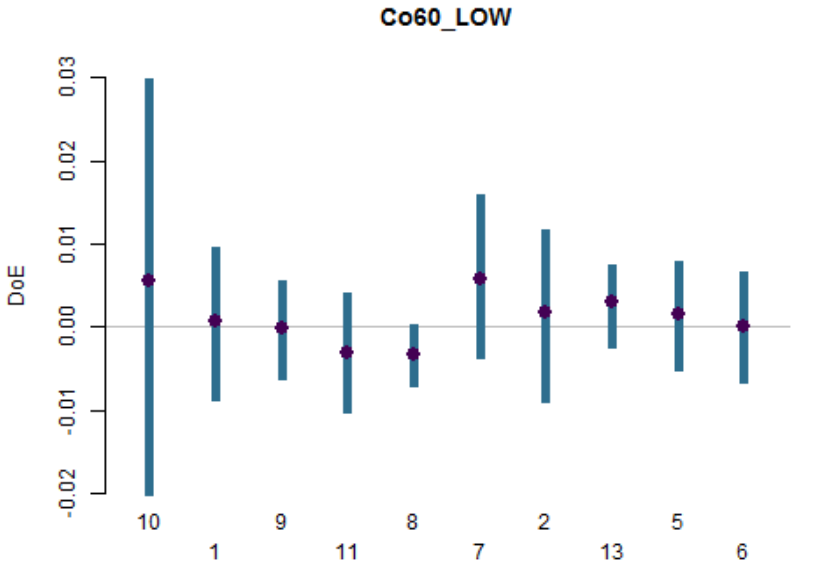


Fig. 17. Low level real concrete: Bayesian estimate of degrees of equivalence for Co-60.

Lab	DoE.x	DoE.U95	DoE.Lwr	DoE.Upr
10	-0.0574	0.0546	-0.1121	-0.0034
1	-0.0198	0.0562	-0.076	0.0365
9	0.0148	0.0517	-0.0366	0.0669
11	0.0257	0.0495	-0.0229	0.0763
8	-0.0068	0.0445	-0.051	0.038
7	0.0154	0.0596	-0.0438	0.0753
2	0.0213	0.0828	-0.0622	0.1033
13	-0.0123	0.0418	-0.0538	0.0297
5	0.0293	0.0348	-0.0068	0.0626
6	-0.0101	0.049	-0.0605	0.0371

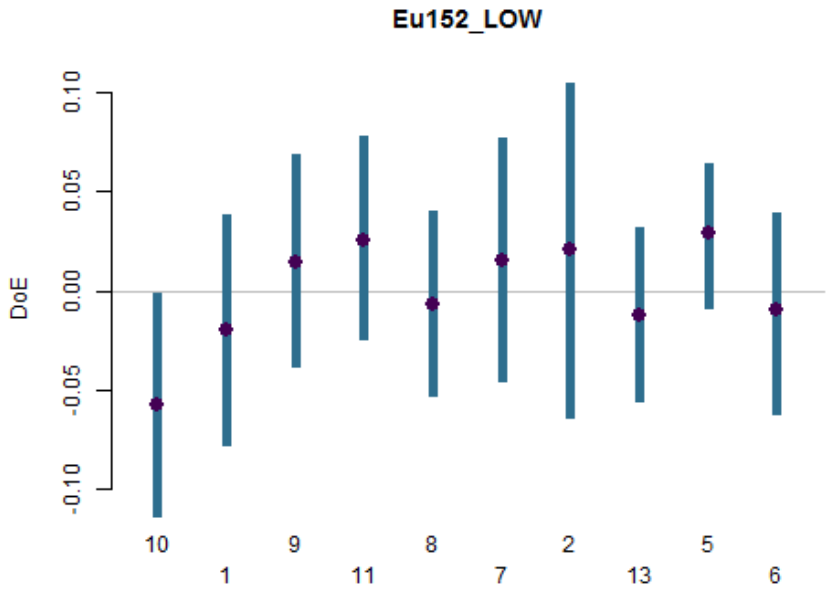


Fig. 18. Low level real concrete: Bayesian estimate of degrees of equivalence for Eu-152.

c. Bayesian estimates for consensus estimates for high level

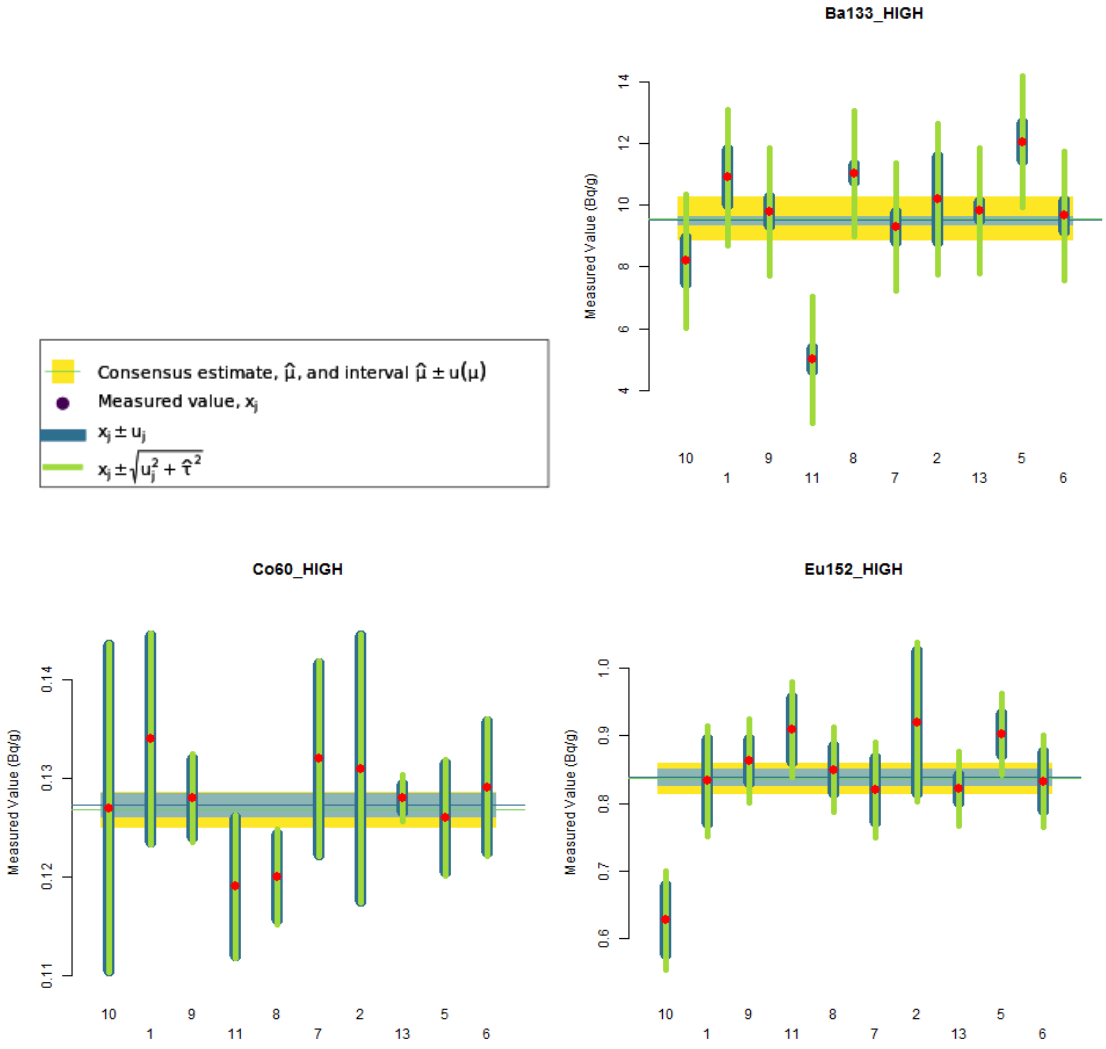


Fig. 19. High level real concrete: Bayesian estimates for Ba-133, Co-60 and Eu-152 mass activity.

Table 10

Results for the consensus estimates using the Bayesian procedure with 100 000 MCMC simulations

	$\hat{\mu}$	$u(\hat{\mu})$	2.5%($\hat{\mu}$)	97.5%($\hat{\mu}$)	$\hat{\tau}$	u($\hat{\tau}$)	2.5%($\hat{\tau}$)	97.5%($\hat{\tau}$)
Ba-133	9.53	0.69	8.16	10.89	2.01	0.53	1.23	3.24
Co-60	0.1269	0.0017	0.1233	0.1301	0.0019	0.0016	1.0010^{-04}	0.0057
Eu-152	0.837	0.022	0.792	0.882	0.051	0.026	0.007	0.109

d. Bayesian estimates for degrees of equivalence for high level

Lab	DoE.x	DoE.U95	DoE.Lwr	DoE.Upr
10	-1.5133	5.0282	-6.6496	3.4008
1	1.4989	4.8926	-3.4342	6.3408
9	0.2501	4.7968	-4.4328	5.1246
11	-5.1121	2.3026	-7.367	-2.7563
8	1.6539	4.457	-2.8873	6.0681
7	-0.2873	4.8695	-4.9424	4.7039
2	0.6943	5.3928	-4.6223	6.1236
13	0.2782	4.5911	-4.4225	4.7939
5	2.7863	4.3973	-1.5644	7.2074
6	0.1146	4.7163	-4.6285	4.7858

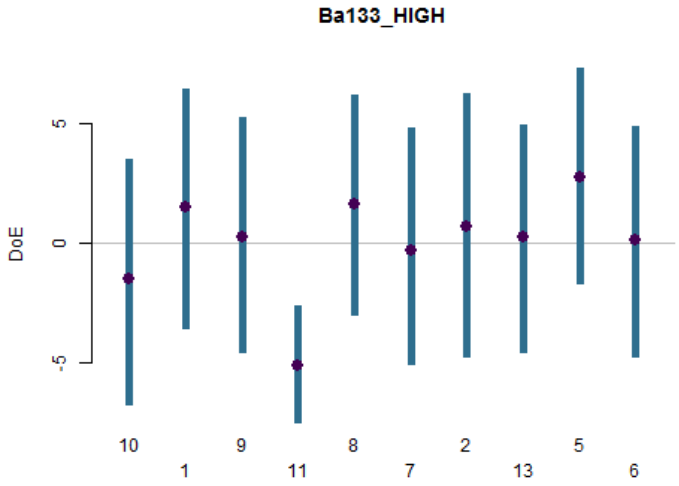


Fig. 20. High level real concrete: Bayesian estimate of degrees of equivalence for Ba-133.

Lab	DoE.x	DoE.U95	DoE.Lwr	DoE.Upr
10	$1.00\ 10^{-04}$	0.0324	-0.0327	0.032
1	0.0073	0.0219	-0.0148	0.0288
9	0.0014	0.0102	-0.0085	0.0119
11	-0.0083	0.015	-0.0232	0.0068
8	-0.0077	0.0103	-0.0181	0.0025
7	0.0053	0.0197	-0.0144	0.0251
2	0.0042	0.0274	-0.0233	0.0315
13	0.0022	0.0083	-0.0061	0.0103
5	-0.001	0.0125	-0.0135	0.0115
6	0.0024	0.0147	-0.0124	0.0169

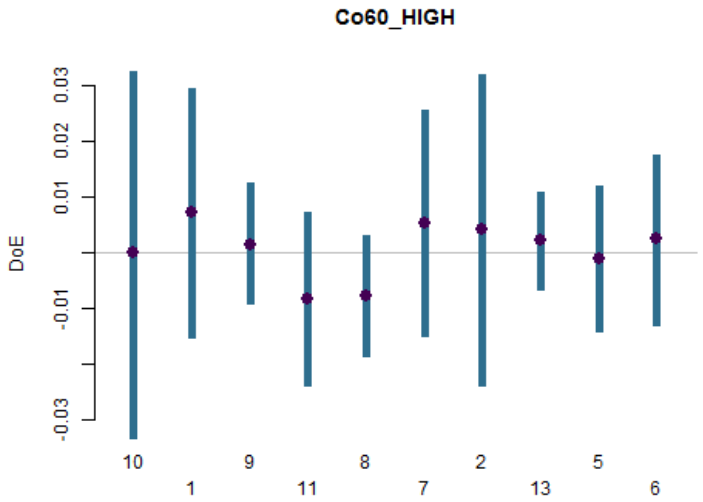


Fig. 21. High level real concrete: Bayesian estimate of degrees of equivalence for Co-60.

Lab	DoE.x	DoE.U95	DoE.Lwr	DoE.Upr
10	-0.2253	0.1158	-0.3412	-0.1102
1	-0.0048	0.1838	-0.1932	0.1763
9	0.0289	0.1553	-0.1256	0.1853
11	0.0786	0.1599	-0.0771	0.242
8	0.0147	0.1547	-0.1379	0.17
7	-0.0189	0.1652	-0.1814	0.1493
2	0.0858	0.246	-0.1578	0.3351
13	-0.0181	0.1528	-0.1672	0.1383
5	0.0744	0.1344	-0.0592	0.2085
6	-0.0051	0.1714	-0.1746	0.1684

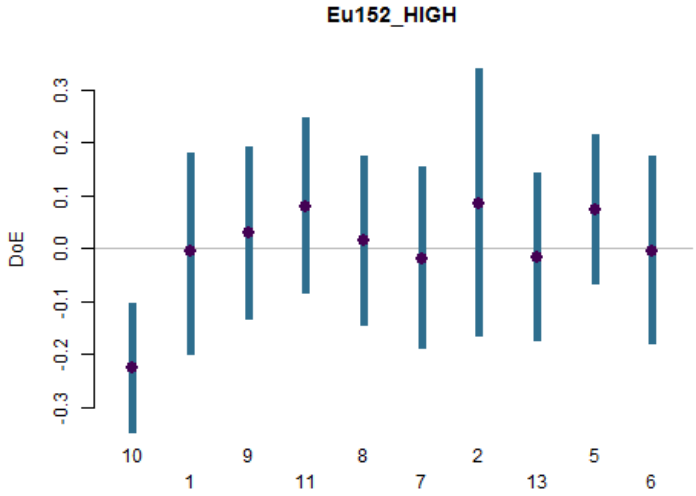


Fig. 22. High level real concrete: Bayesian estimate of degrees of equivalence for Eu-152.