 https://doi.org/10.3011/ESARDA.IJNSNP.2018.10

Approximate Bayesian Computation Applied to Nuclear Safeguards Metrology

Tom Burr[1], Thomas Krieger[2], Claude Norman[1]

[1] Safeguards Information Management, IAEA, Vienna Austria
[2] Forschungszentrum Jülich, Inst. Energy and Climate Research, Jülich, Germany
E-mail: t.burr@iaea.org, t.krieger@fz-juelich.de, c.norman@iaea.org

Abstract:

Approximate Bayesian Computation (ABC) is an inference option if a likelihood for measurement data is not available, but a forward model is available that outputs predicted observables, such as gamma counts, for any set of specified input parameters, such as item mass. This paper reviews ABC and illustrates how ABC can be applied in safeguards metrology. A key aspect of metrology is uncertainty quantification (UQ), approached from physical first principles ("bottom-up") or approached empirically by comparing measurements from different methods and/or laboratories ("top-down"). Although ABC is not yet commonly used in metrology, an example using enrichment measurements is used to illustrate potential advantages in ABC compared to current bottom-up approaches. Using the same example, ABC is also shown to be useful in top-down UQ. And, the example shows good agreement between bottom-up and top-down measurement error relative standard deviation (RSD) estimates, while also allowing for the effects of item-specific biases. As a diagnostic, in applications of ABC, the actual coverages of probability intervals are compared to the true coverages. For example, if an ABC-based interval for the true measurement RSD is constructed to contain approximately 95% of the true values, then one can check whether the actual coverage is close to 95%. It is shown that one advantage of ABC compared to other Bayesian approaches is its apparent robustness to miss-specifying the model while maintaining good agreement between the nominal and the actual coverage.

Keywords: approximate Bayesian computation; metrology; non-destructive assay; uncertainty quantification

1. Introduction

Nuclear safeguards aim to verify that nuclear materials are used exclusively for peaceful purposes. To ensure that States honor their safeguards obligations, measurements of nuclear material inventories and flows are needed. Statistical analyses used to support conclusions require UQ, usually by estimating the RSD in random and systematic errors associated with each measurement method [1-9].

To monitor for possible data falsification by the operator that could mask nuclear material diversion, paired (operator, inspector) data are assessed. These paired data are declarations usually based on measurements by the operator, often using destructive assay, and measurements by the inspector, often using non-destructive assay (NDA). Statistical tests are applied one-item-at-a-time, and also to assess for a possible trend by computing the overall difference of the operator-inspector values using the

D statistic, one version of which is defined as $D = \frac{N}{n}\sum_{j=1}^{n}\frac{O_j - I_j}{O_j}$ where j indexes the sample items, O_j is the operator declaration, I_j is the inspector measurement, n is the verification sample size, and N is the total number of items in the stratum. The D statistic and the one-item-at-a-time tests rely on estimates of operator and inspector measurement error RSDs that are based on top-down UQ from previous inspections [1,2]. Inspector NDA measurements are made using portable neutron and gamma detectors taken into the facility, which involves challenges for UQ (Section 3). Such an assessment depends on the assumed measurement error model (for example, if the errors scale with the true value then a relative error model is appropriate) and associated uncertainty components, so it is important to perform effective UQ [2,3,4,8,9].

Another quantitative assessment in safeguards that requires UQ involves the material balance defined as $MB = T_{in} + I_{Begin} - T_{out} - I_{end}$, where T is transfers and I is inventory. The covariance Σ_{MB} of a sequence of n material balances is an n-by-n matrix with the MB variances on the diagonal and the covariances between pairs of MBs on the off-diagonals. The entries in Σ_{MB} are estimated using measurement error variance propagation applied to estimates of the RSDs in random and systematic error variances for each of the operator's measurement methods [1, 4-7].

MB evaluations and verification data assessments rely on estimates of measurement error RSDs. Historical paired (operator, inspector) data is used for top-down UQ, applying analysis of variance (ANOVA), to estimate RSDs. Bottom-up UQ propagates errors in all key steps of the assay to predict the RSD in the estimated nuclear material mass; this error propagation is similar to that used in the guide to expression of uncertainty in measurements [GUM, 10]. It is common for RSD estimates from bottom-up UQ to be smaller than those from top-down UQ [2, 9]. Currently,

Burr, T., Krieger, T., & Norman, C. (2018). Approximate Bayesian Computation Applied to Nuclear Safeguards Metrology. *ESARDA Bulletin - The International Journal of Nuclear Safeguards and Non-proliferation*, 57, 50-59. https://doi.org/10.3011/ ESARDA.IJNSNP.2018.10

a gap between bottom-up and top-down RSD estimates does not directly impact inspectors' conclusions, because the top-down RSD estimates are used to set alarm thresholds in MB evaluations and verification data assessments. However, only when there is good agreement between bottom-up and top-down UQ can the potential to improve an NDA method be fully understood.

Because MB evaluations and verification data assessments rely on top-down estimates of random and systematic (see Section 2) measurement errors, top-down RSD estimates set a target for bottom-up UQ. One step to improve UQ is to improve bottom-up UQ so that its RSD estimates are in better agreement those from top-down UQ [2, 9]. Another step to improve UQ is to estimate the uncertainty in the RSD estimates so that any gap between bottom-up and top-down RSD estimates can be assessed for significance (which is not currently done in practice). Toward the goal of improving UQ, this paper introduces ABC for both bottom-up and top-down UQ. Any Bayesian approach provides a probability distribution for the unknown model parameters, which are the unknown random and systematic RSDs in this context, so the uncertainties in the RSD estimates are known. And, ABC has two potential advantages over other Bayesian methods in this context. First, ABC appears to be more robust to small or modest misspecifications of the data likelihood. Second, ABC can easily accommodate comprehensive bottom-up UQ, including effects such as uncertainties in nuclear data and model-based adjustment of test items to calibration items [9].

This paper is organized as follows. Sections 2 and 3 describe top-down and bottom-up UQ, respectively. Section 4 describes approximate Bayesian computation (ABC [11-13]). Section 5 applies ABC to top-down and bottom-up UQ for safeguards for NDA using the enrichment meter principle (EMP [14-16]). Section 6 is a summary.

2. Top-down UQ applied to paired (operator, inspector) data

An effective measurement error model must account for variation within and between groups, where a group is, for example, a calibration or inspection period. A typical model for relative errors for the inspector (*I*) (and similarly for the operator *O*) is

$$I_{jk} = \mu_{jk}(1 + S_{Ij} + R_{Ijk}), \quad (1)$$

where I_{jk} is the inspector's measured value of item k (from 1 to n) in group j (from 1 to g), is the true but unknown value of item k from group j, $R_{Ijk} \sim N\left(0, \delta_{RI}^2\right)$ is a random error of item k from group j, $S_{Ii} \sim N\left(0, \delta_{SI}^2\right)$ is a short-term systematic error in group j. To better understand Eq. (1), Fig. 1 plots 10 simulated values in each of 3 groups of $d = (O - I) / O$ values. Section 5.1 contains more information regarding Fig. 1.

The measurement error model sets the stage for applying ANOVA with random effects [17-19]. Neither $R_{I_{ij}}$ nor S_{I_i} are observable. However, for various types of observed data, one can estimate their respective variances δ_{RI}^2 and δ_{SI}^2. For the error model in Eq. (1), the standard deviation σ_D of D, is $\sigma_D = \sqrt{N^2\left(\frac{\delta_R^2}{ng} + \frac{\delta_S^2}{g}\right)}$ where $\delta_R^2 = \delta_{RO}^2 + \delta_{RI}^2$ and $\delta_S^2 = \delta_{SO}^2 + \delta_{SI}^2$, so alarm thresholds for D that correspond to user-specified false alarm probabilities can be selected. Similarly, the one-at-a-time tests also require estimates of δ_R^2 and δ_S^2, which are obtained by applying random one-way ANOVA to real paired difference data that are assumed to follow Eq. (1). Reference [3] evaluates impacts on alarm probabilities of using estimates of δ_R^2 and δ_S^2 instead of the true quantities In some safeguards contexts such as MB evaluation, the estimates of δ_R^2 and δ_S^2 must be partitioned into δ_{RO}^2 *and* δ_{RI}^2 and δ_{SO}^2 *and* δ_{SI}^2, respectively [2, 3]. Note from the expression for σ_D that δ_R^2 is divided by the number of observations ng, and that δ_S^2 is divided by the number of periods g, which makes sense according to the error model (1) and in view of Figure 1.

Error model (1) does not include long-term systematic error. The short-term systematic error is assumed to change between inspection periods [14,19] due to re-calibration and possibly other effects. In practice, there are sometimes tests for long-term systematic error, where long-term means as long as (or longer than) the data evaluation period, which is typically multiple inspection periods or years. Any long-term error is investigated and will be assumed in this paper to be zero.

3. Bottom-up UQ

NDA uses calibration and/or modelling to infer nuclear material (NM) mass using detected radiation such as neutron and gamma emissions. Three issues in UQ for NDA are:

1. NDA is applied in challenging settings because the detector is brought to the facility where ambient conditions can vary over time, and the items are often heterogeneous in some way. Because of such challenges, dark uncertainty [20] can be large, as is evident whenever bottom-up UQ predicts smaller RSD than is observed in top-down UQ.
2. There is no UQ guide for NDA that is analogous to the GUM. But, the GUM is typically followed for the error variance propagation steps in UQ, and each NDA method has a specific and documented implementation of UQ (for example, ASTM C1514 [15] for the EMP).
3. NDA is often used when test items differ substantially from calibration items; therefore, the concept of item-specific bias is important, and is addressed in Section 5.

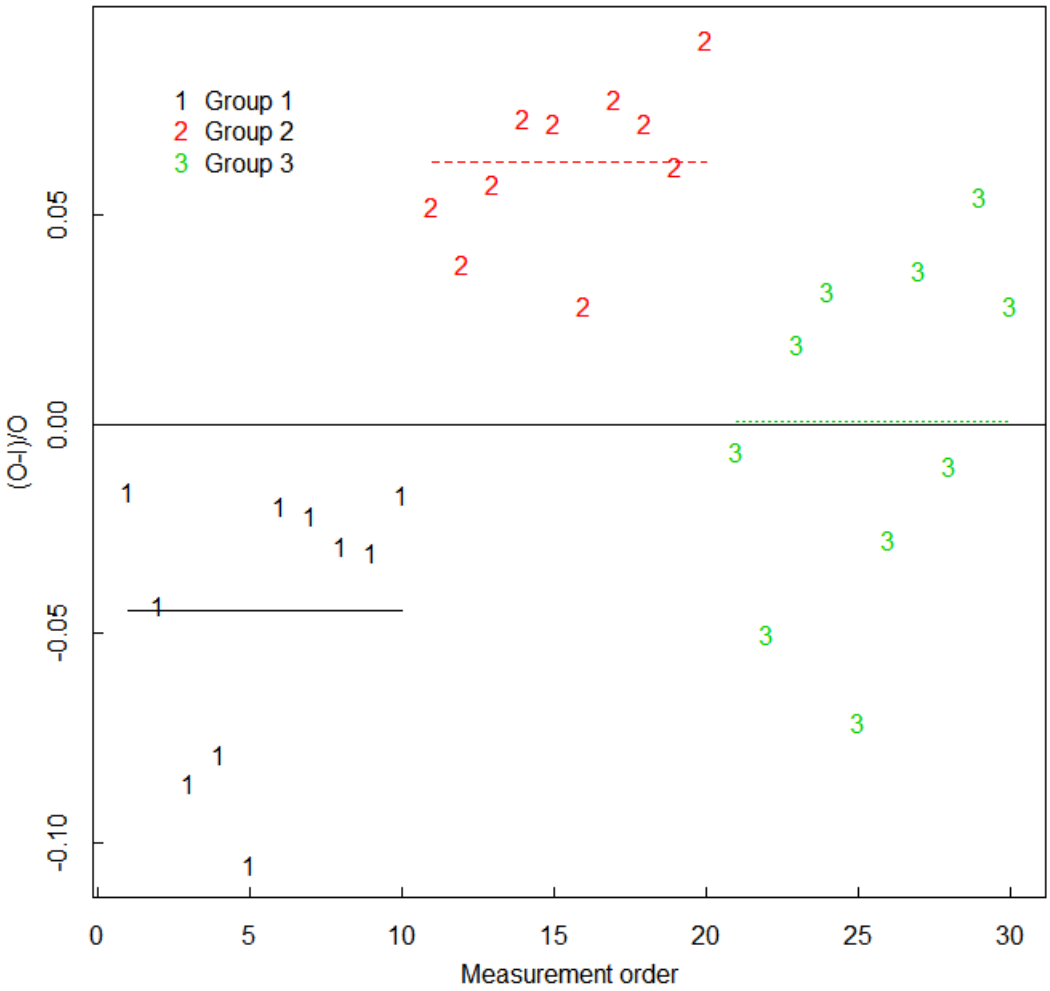

Figure 1: Example (simulated) of 10 $d = (O - I) / O$ values in each of 3 groups.

In NDA, error variance propagation is used as a component of bottom-up UQ by propagating errors in inputs. Bottom-up UQ is often approached by using the GUM's measurement equation, expressed as

$$Y = f(X_1, X_2, \ldots, X_N) \tag{2}$$

for measurand Y and inputs $X_1, X_2, \ldots, X_N$. The GUM applies the delta method to Eq. (2) to propagate error variances in the X_i to estimate the standard deviation in Y. The input quantities can include, for example, measured count rates, estimates of calibration parameters or other measurands, such as measured values in steps an assay method. The delta-method assumes that $f(X_1, X_2, \ldots, X_N)$ in Eq. (2) can be well approximated by a first-order Taylor series expansion around the mean values of each, and then the linear approximation to $f(X_1, X_2, \ldots, X_N)$ can be used to estimate σ_Y^2 given estimates of the variances for each X_i (and, correlations between the can be accommodated). If the first-order Taylor series is not sufficiently adequate, the GUM recommends Monte Carlo simulation. Note that Eq. (2) implies that Y is random, so the GUM implicitly adopts a Bayesian viewpoint (Section 4) without explicitly stating a prior distribution for Y [21, 22].

Recently, the NDA community is recognizing a need for more comprehensive bottom-up UQ that thoroughly addresses uncertainty in model-based adjustments of test items to calibration items [2,9]. Toward that goal, several US national laboratories are collaborating on a multi-year project to improve UQ for NDA and the standard committee ASTM C26.12 is another group also working on UQ for NDA. One possible outcome of these collaborations is better guidance on bottom-up UQ for calibration data that allows for both errors in predictors and for item-specific bias. It is also possible that approaches for better bottom-up UQ will be provided in the next version of the GUM [21, 22].

4. ABC

Bayesian ANOVA such as could be applied to data generated from Eq. (1) has been studied [17], and Bayesian methods are slowly being adopted in metrology [9,10,21,22]. However, Bayesian ANOVA using ABC has not been well studied. In any Bayesian approach, prior information

regarding the magnitudes and/or relative magnitudes of δ_{RI}^2 and δ_{SI}^2 can be provided [21-23]. If the prior is "conjugate" for the likelihood, then the posterior is in the same likelihood family as the prior, in which case analytical methods are available to compute posterior prediction intervals for quantities of interest. In order that a wide variety of priors and likelihoods can be accommodated, modern Bayesian methods do not rely on conjugate priors, but use numerical methods to obtain samples of δ_{RI}^2 and δ_{SI}^2 from their approximate posterior distributions [23]. For numerical methods such as Markov Chain Monte Carlo [23], the user specifies a prior distribution for δ_{RI}^2 andδ_{SI}^2, and a likelihood (which need not be normal). ABC does not require a likelihood for the data (but this section provides clarification regarding the need for a likelihood in this NDA context), and, as in any Bayesian approach, ABC accommodates constraints on variances through prior distributions [11-13, 24-26].

The "output" of any Bayesian analysis is the posterior distribution for each model parameter, and so the output of ABC for data generated from Eq. (1) is an estimate of the posterior distributions of δ_{RI}^2 and δ_{SI}^2. No matter what type of Bayesian approach is used, a well-calibrated Bayesian approach satisfies several requirements. One requirement is that in repeated applications of ABC, approximately 95% of the middle 95% of the posterior distribution for each of δ_{RI}^2 and δ_{SI}^2 should contain the respective true values. That is, the actual coverage should be closely approximated by the nominal coverage. A second requirement is that the true standard deviation of the ABC-based estimates of δ_{RI}^2 and δ_{SI}^2 should be closely approximated by the standard deviation of the ABC-based posterior distributions of δ_{RI}^2 and δ_{SI}^2. Inference using ABC can be briefly summarized as follows:

ABC can be described using this high-level algorithm description:

ABC Inference

For i in 1, 2, …, N

1. Sample θ from the prior, $\theta \sim f_{prior}(\theta)$.
2. Simulate data y' from the model $y' \sim P(y \mid \theta)$.
3. Denote the real data as y. If distance $d(S(y'),S(y)) \le \varepsilon$, *accept* θ *as an observation from* $f_{posterior}(\theta \mid y)$.

Experience with ABC suggests that the ABC approximation to $f_{posterior}(\theta \mid y)$ improves if step (3) is modified to include a weighting function so that values of $\theta \sim f_{prior}(\theta)$ that lead to very small values of the distance $d(S(y'),S(y))$ are weighted more heavily in the estimated posterior [24,25].

In ABC, the model has input parameters θ and outputs data $y(\theta)$ and there is corresponding real data y_{obs}. For example, the model could be Eq. (1), which specifies how to generate synthetic *I* (or *O*) data, and does require a likelihood; however, the true likelihood used to generate the data need not be known to the user. Synthetic data is generated from the model for many trial values of θ, and trial θ values are accepted as contributing to the estimated posterior distribution for $\theta|\ y_{obs}$ if the distance $d(y_{obs},\ y(\theta))$ between y_{obs} and $y(\theta)$ is reasonably small. Alternatively, for most applications, it is necessary to reduce the dimension of y_{obs} to a small set of summary statistics $S(y_{obs})$ and accept trial values of θ if $d(S(y_{obs}),S(y(\theta)) < \varepsilon$, where ε is a user-chosen small threshold near 0. Here, for example, $y_{obs} = d = \frac{O-I}{O}$ data in each inspection group, and $S(y_{obs})$ includes within and between groups sums of squares Specifically, the ANOVA-based estimator of δ_{RI}^2 in Eq. (1) is $\hat{\delta}_R^2 = \frac{1}{n-g}\{\sum_{j=1}^{g}\sum_{k=1}^{n}(d_{jk}-\bar{d}_j)^2\}$,and the usual estimate of δ_{SI}^2 is $\hat{\delta}_S^2 = \frac{\sum_{j=1}^{g}(\bar{d}_j-\bar{\bar{d}})^2}{g-1} - \frac{\hat{\delta}_R^2}{n}$. The quantities $\hat{\delta}_R^2$ and $\hat{\delta}_S^2$ are therefore good choices for summary statistics for ABC. Recall that because trial values of θ are accepted if $d(S(y_{obs}),S(y(\theta)) < \varepsilon$, an approximation error to the posterior distribution arises that several ABC options attempt to mitigate. Recall also that such options weight the accepted θ values by the actual distance $d(S(y_{obs}),S(y(\theta))$ (abctools [25] in R [26]).

To summarize, ABC applied to data following Eq. (1) consists of three steps: (1) sample parameter values of δ_R^2 and δ_S^2 from their prior distribution $p_{\text{prior}}(\theta)$; (2) for each simulated value of θ in (1),simulate data from Eq. (1); (3) accept a fraction of the sampled prior values in (1) by checking whether the summary statistics computed from the data in (2) satisfy $d(S(y_{obs}),S(y(\theta))\ < \varepsilon$. If desired, aiming to improve the approximation to the posterior, adjust the accepted θ values on the basis of the actual $d(S(y_{obs}),S(y(\theta))$ value. ABC requires the user to make three choices: the summary statistics, the threshold ε, and the measure of distance *d*. Reference [11] introduced a method to choose summary statistics that uses the estimated posterior means of the parameters based on pilot simulation runs. Reference [12] used an estimate of the change in posterior $p_{\text{posterior}}(\theta)$ when a candidate summary statistic is added to the current set of summary statistics. Reference [13] illustrated a method to evaluate whether a candidate set of summary statistics leads to a well-calibrated posterior, in the same sense that is used in this paper; that is, nominal posterior probability intervals should

have approximately the same actual coverage probability, and the posterior variance should agree with the observed variance in testing.

5. EMP Example

The mass of ^{235}U in an item can be estimated by using the measured net weight of uranium U in the item and the measured ^{235}U enrichment (the ratio ^{235}U/U). Enrichment can be measured using the 185.7 keV gamma-rays emitted from ^{235}U by applying the EMP. The EMP aims to infer the enrichment by measuring the count rate of the strongest-intensity direct (full-energy) gamma from decay of ^{235}U, which is emitted at 185.7 keV [14-16]. The EMP assumes that the detector field of view into each item is identical to that in the calibration items (the "infinite thickness" assumption), that the item is homogeneous with respect to both the ^{235}U enrichment and chemical composition, and that the container attenuation of gamma-rays is the same as or similar to that in the calibration items so that empirical correction factors have modest impact and are reasonably effective. If these three assumptions are met, the known physics implies that the enrichment of ^{235}U in the U is directly proportional to the count rate of the 185.7 keV gamma-rays emitted from the item. It has been shown empirically that under good measurement conditions, the EMP can have a random error RSD of less than 0.5 % and a long term bias of less than 1 %, depending on the detector resolution, stability, and extent of corrections needed to adjust items to calibration conditions. Some bottom-up UQ examples for the EMP in [14,16,19] have estimated random error RSD ranging from less than its 0.5% target to approximately 1.0% (because of item-specific biases arising due to container thickness variations and other effects,) but less than the 2% to 4% reported from corresponding top-down UQ for the ^{235}U mass in UO_2 drums. Also, top-down UQ reports total error RSD (random and short-term systematic) of 4% to 20 % for some items analyzed in [19] (the RSD tends to be larger for smaller values of enrichment).

The known nominal enrichment in each of several standards can be fit to observed counts in a few energy channels near the 185.7 keV energy as the "peak" region and to the counts in a few nearby energy channels below and above the 185.7 keV energy but outside the peak area to estimate background (two-region EMP method), expressed as

$$Y = \beta_1 N + R_Y \qquad (3),$$

where Y is the enrichment, N is the peak count rate near 185.7keV, R_Y is random error and β_1 is a calibration constant. Figure 2 is an example low-resolution (NaI detector) gamma spectrum near the 185.6keV region. The gross count and the two background ROI counts can be combined into one net count, resulting in one predictor as in Eq. (3). For example, if the same number of energy channels are used for both the peak and background ROI, then Net count rate = Peak count rate – Background count rate. There is usually non-negligible error in N, so errors in predictors cannot be ignored when fitting Eq. (3) to calibration data [14]. Alternatively, both peak and background counts can be used as predictors [14-16]. There will be measurement errors in the gross and background count rates and there will often be correction factors applied, for example, to adjust test item container thickness to calibration item container thickness. There is much literature regarding errors in predictors and whether to fit Y as a function of N (reverse calibration) or to fit N as a function of Y and invert to solve for Y (inverse calibration). Both options should be investigated using simulation, because analytical approximations have been shown to not be sufficiently accurate either to decide between options or to assess the uncertainty in the chosen option [14,27]. However, the root mean squared prediction error (RMSE) of reverse calibration (Eq. (3) is an example of reverse calibration) has been generally found to be the same as or smaller than that of inverse calibration.

Calibration data is used to compute the estimate $\hat{\beta}_1$ of the model parameter β_1 in Eq. (3). The variance of $\hat{\beta}_1$ is not necessarily well approximated by the usual least squares expression because of errors in N. Therefore, [14,27] suggest that the RMSE in $\hat{Y}$ be estimated by simulating the calibration procedure, which allows for errors in N arising from Poisson counting statistics, and also arising from other sources, such as container thickness (with or without an adjustment for the measured container thickness) varying among test items. Errors in N due to imperfect adjustment for container thickness can manifest as item-specific bias. The ABC strategy below illustrates how item-specific bias can be understood and estimated. The RMSE in $\hat{Y}$ is defined as usual, as $E((\hat{Y} - Y_{true})^2) = E(\hat{Y} - E(\hat{Y})^2) + (E\hat{Y} - Y_{true})^2 =$ = variance + bias^2.

Note that one can express the calibration Eq. (3) as in Eq. (2), where X_1 is $\hat{\beta}_1$ and X_2 is N, with var($\hat{\beta}_1$) estimated by simulation, so GUM's Eq. (2) could be used to estimate var($\hat{Y}_1$) and cov($\hat{Y}_1$, $\hat{Y}_2$) , although [22] points out that GUM's Eq. (2) is not actually designed to be applied to calibration applications, regardless of whether there are errors in the predictors.

In general, item-specific bias can arise due to item-specific effects, expressed as

$$CR/M = g(X_1, X_2, \ldots, X_N), \qquad (4),$$

where CR is the item's neutron or gamma count rate, M is the item NM mass, g is a known function, and $X_1, X_2, \ldots, X_N$ are N auxiliary predictor variables such as item density, source NM heterogeneity, and container thickness, which will generally be estimated or measured with error and so are regarded as random variables . To map Eq. (4), to GUM's Eq. (2), write

$$M = CR \,/\, g(X_1, X_2, \ldots, X_N) = h(X_1, X_2, \ldots, X_M) \qquad (5),$$

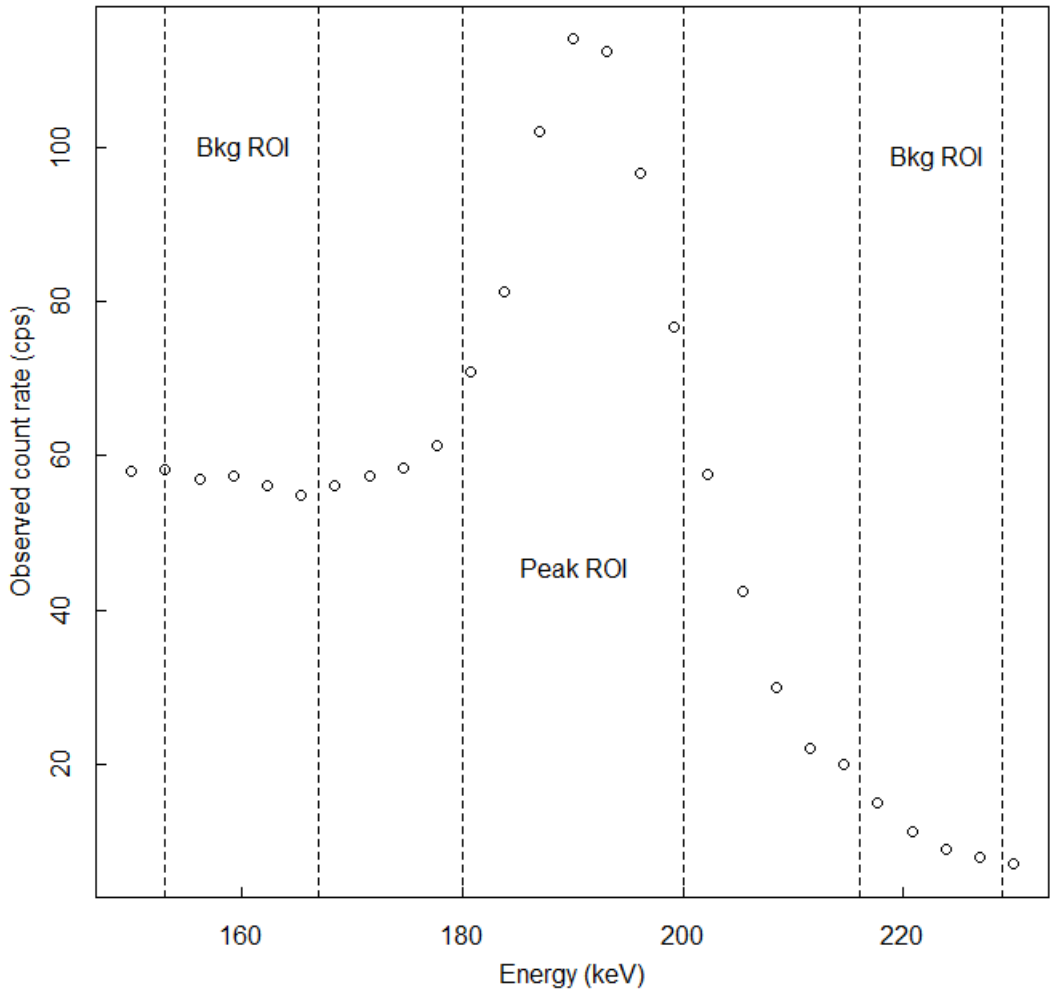

Figure 2: Example low-resolution (NaI detector) gamma spectrum near the 185.6keV peak with two background regions (one region below the 185.7 keV peak and one region above the 185.7 keV peak).

where the measured CR is now among the $M = N+1$ inputs. Note that Eq. (5) is the same as Eq. (4), but some of the X_i account for item-specific departures from reference items used for calibration. More specifically, Eq. (3) can be re-expressed as

$$Y = \beta_1(item)N + R_Y \quad (6),$$

where the calibration constant $\beta_1(item)$ varies across items and R_Y is the random error in Y. Equation (6) is a random-coefficient regression equation, and real and/or simulated data generated from Eq. (6) can be used to estimate the average value of $\beta_1(item)$. Eq. (6) is a model that can explain item-specific bias, which is usually regarded as a random error (across items). Many NDA examples adjust test items to calibration items using some type of modelling [2,14]. In the EMP, an additional input variable X_3 could be an adjustment for container thickness to be applied to the detected net count rate in Eq. (6). And, one way to model the effect of imperfect adjustment for each item's container thickness is to include another random error in simulated net count rates used as synthetic calibration data, rather than to modify β_1. In practice, net count rates are sometimes adjusted to account for the measured container thickness, using Beer's law, which states that the gamma intensity after passing through a container with density ρ, attenuation coefficient μ and thickness t is multiplied by $\exp(-\mu\rho t)$). Note that errors in N have the same impact as errors in $\beta_1(item)$ because the term $\beta_1(item)N$ appears in Eq. (6).

5.1 ABC applied to the EMP

The purpose of this bottom-up example is to show how to apply ABC and to show how ABC makes a bottom-up estimate of random and systematic RSDs such as those illustrated in Figure 1, and how ABC includes uncertainty in the estimated RSDs. ABC applied to the EMP can be implemented in the following 7 steps.

1) Estimate the average regression coefficient $\hat{\beta}_1$ in Eq. (6) using available real calibration data, typically consisting of approximately 3 to 5 (Y,N) pairs. The real calibration data

used here are Y = 0.355, 0.80, 2.175, 3.305, 5.0 (^{235}U enrichments of 5 standards) and the corresponding N= 0.062, 0.139, 0.37, 0.575, 0.866 net count rates.

2) Use the estimate $\hat{\beta}_1$ from (1) to generate many ($S = 10^5$ or more) synthetic calibration runs using $Y = \beta_1(item)N + R_Y$ to generate synthetic sets of 5 paired (Y,N) values, with run i producing the estimate $\hat{\beta}_{1,i}$. This example generated the $\beta_1(item)$ values randomly and uniformly from 0.85 to 0.95.
3) Specify a prior distribution for the true enrichment μ_Y. If little is known about the true enrichment values, then, for example, specify a uniform prior ranging from the lowest possible true enrichment to the highest possible true enrichment. This example used wide a uniform distribution from 0.355 to 5.0, which avoids extrapolating outside the range of the true enrichments.

4) Specify a background count rate μ_B (this example used μ_B = 0.05) and use the estimated regression coefficient $\hat{\alpha}_1$ from the regression equation $N = \alpha_1(item)Y + R_N$ to generate a net count rate μ_N that corresponds to a value of μ_Y sampled from its prior distribution. This example used an RSD in Y of 0.1% and in R_N of 5%.

5) Specify a count time (this example used 600 seconds) t, simulate $B \sim Poisson(\mu_B t)$, $G \sim Poisson(\mu_G t)$, and compute a net count rate (assuming the same number of energy channels for the peak and background ROIs) $N = \frac{G}{t} - \frac{B}{t}$.

6) Repeat (4) and (5) many (10^5 or more) times to construct a large collection of simulated true enrichments μ_Y and corresponding net count rates N, which is an effective summary statistic.

(7) For each simulated test case, simulate a value of μ_Y from its prior, use steps (4) and (5) to generate N_{test}, and compute the distance $d(N_{test}, N_i) = |N_{test} - N_i|$ from N_{test} to each of the i = 1, 2, …, 10^5 realizations from step (6), and accept those μ_Y generated in step (6) that correspond to $|N_{test} - N_i| \le \varepsilon$ as observations from the posterior $\mu_Y \mid N$ (which in this case is somewhat complicated to specify analytically) weighting inversely by the distance $|N_{test} - N_i|$ if desired. Linear regression was not used in this ABC implementation for predicting μ_Y for each simulated test value of N, although it could have been, and note that regression is used in step (2) to generate the 10^5 pairs of (μ_Y,N) in the training data for ABC.

The result in applying steps 1-7 is an estimate of the posterior distribution for the true enrichment μ_Y, similar to that in Fig. 3, as explained below. To assess ABC performance, the two criteria mentioned can both be used: the estimated standard deviation of the posterior should be in good agreement with the observed standard deviation across test items, and the nominal probability interval coverage should also be in good agreement with the actual coverage. The data plotted in Fig. 1 were generated using the steps just given to apply ABC for both operator and inspector data, assuming for simplicity that both used the EMP and both recalibrated at the beginning of periods 1, 2, and 3. The estimated standard deviation of $d_{rel} = (O - I)/O$ (which includes both within- and between-group standard deviations) from top-down data such as that in Fig. 1 (also using ABC as outlined in Section 4) is 0.11, which is very close to that predicted from the bottom-up ABC (0.12 as explained in the next paragraph) posterior standard deviations for O and I.

Recall from Section 4 that the usual ANOVA-based estimator of σ^2_{Rd} (using the multiplicative form of Eq. (1) for both operator and inspector) is $\hat{\sigma}^2_{Rd} = \frac{1}{n-g}\{\sum_{j=1}^{g}\sum_{k=1}^{n}(d_{jk} - \bar{d}_j)^2\}$, and the usual estimate of σ^2_{Sd} is $\hat{\sigma}^2_{Sd} = \frac{\sum_{j=1}^{g}(\bar{d}_j - \bar{\bar{d}})^2}{g-1} - \frac{\hat{\sigma}^2_R}{n}$. The quantities, $\hat{\sigma}^2_{Rd}$ and $\hat{\sigma}^2_{Sd}$ are therefore good summary statistics for ABC, and were used to implement ABC for the top-down analysis of data such as that in Fig. 1.

The 0.12 bottom-up prediction for the standard deviation of $d_{rel} = (O - I)/O$ is illustrated by plotting the posterior for O for a particular N value in Fig. 3, which has a total (random plus systematic) RSD of 0.08 (from the 7-step procedure). Because this example assumes both O and I made the same type of EMP measurements, the bottom-up prediction of the RSD for $d_{rel} = (O - I)/O$ is given by $\sqrt{(0.08^2 + 0.08^2)} = 0.11$ (from bottom-up). The 0.12 top-down estimate of the RSD of $\delta_{d_{rel}}$ (see Fig. 4, using data such as the data in Fig. 1) is the RSD of the ABC-based posterior distribution for $\delta_{d_{rel}}$ from top-down UQ, with g = 3 groups and n = 10 paired measurements per group (as in Fig.1). The 0.12 estimate has an associated 14% RSD, and an approximate 95% probability interval for $\delta_{d_{rel}}$ is 0.086 to 0.15.

One advantage of having a probability interval for both the bottom-up and top-down estimate of $\delta_{d_{rel}}$ is that one can assess whether differences between the top-down and bottom-up estimates of $\delta_{d_{rel}}$ are significant. In this example, bottom-up UQ using ABC agrees very well with corresponding top-down UQ using ABC that used simulated O and I values as in Fig. 1; which means that in this application, ABC is well-calibrated. Trial and error was used to select $\varepsilon = 0.01$ to obtain good agreement between the ABC-based predicted standard deviation and the observed standard deviation. Coverages of the ABC-based probability intervals were checked and, as mentioned, excellent agreement between nominal and actual was observed. Specifically, the 99%, 95%, and 90% probability intervals contained approximately 99%, 95%, and 90%, respectively of the true values of μ_Y.

Because bottom-up RSD estimates are often compared to top-down RSD estimates to look for un-modelled effects ("dark uncertainty" [20]), it is important for RSD estimates to include information regarding uncertainty in the estimated RSDs. In this example, ABC provides estimates of the uncertainty in the parameter estimates (in this case, the estimated RSDs) in the same manner that any Bayesian analysis does, by providing a posterior distribution for each parameter. Because the top-down and bottom-up RSD estimates are essentially the same in this example, there is no evidence of dark uncertainty (and there should not be, because no dark uncertainty was simulated).

Assuming a normal distribution is not always a good approximation for the actual distribution of $(O\text{-}I)/O$ values used in top-down UQ. So, regarding robustness of ABC in top-down UQ, it has been found that the actual coverages are essentially the same (to within simulation uncertainty) as the nominal coverages, at 90%, 95%, and 99% probabilities, for a normal distribution and all of the non-normal distributions investigated (uniform, gamma, lognormal, beta, *t*, and generalized lambda with thick or thin tails) for the distribution of the random error term R_Y in Eq. (6). Regarding robustness of ABC in the bottom-up context, a key aspect of ABC is the ease with which different forward models linking model parameters (such as the true RSDs in Eq. (2)) to model output and corresponding summary statistics. For example, the Poisson model used in the ABC implementation for the EMP can be easily replaced with an overdispersed Poisson model if exploratory analysis of real data suggests overdispersion.

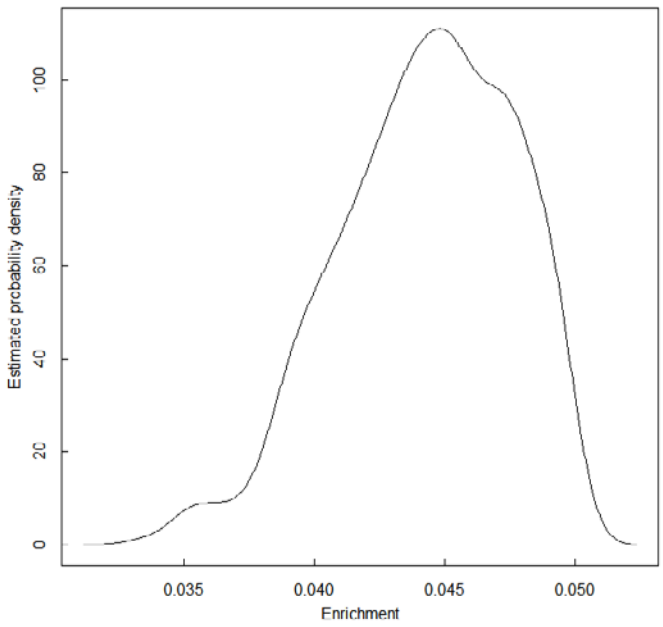

Figure 3: The bottom-up ABC-based estimate of the posterior δ_{TI} (or δ_{TO}).

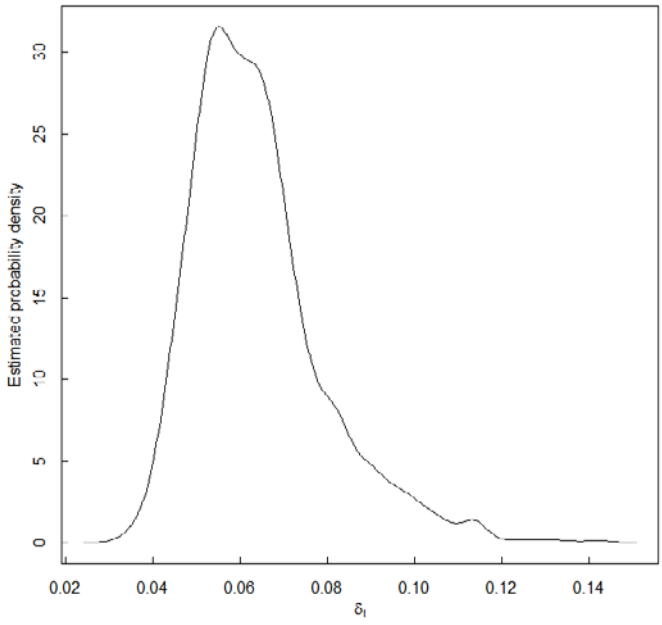

Figure 4: The top-down ABC-based estimate of the posterior for δ_T with RSD of 14%.

6. Discussion and Summary

ABC was used for both bottom-up and top-down RSD estimation in simulated EMP data (using a calibration set of 5 real EMP data pairs). ABC provided robust estimates of the posteriors for model parameters (the RSD values), so bottom-up RSD estimates could be compared to top-down estimates while accounting for parameter uncertainty (as defined by the width of the posterior).

ABC is very well-suited for bottom-up UQ in more challenging NDA applications, for example, when the measurement data is summarized using higher-dimensional summary statistics, such as the estimated net areas in peak regions of interest in gamma spectrometry [28,29], using microcalorimetry. Current microcalorimetry algorithms fit approximately 15 peak areas (associated with gamma ray energies) associated with different isotopes of Pu, U, and Am. These 15 peak areas are the summary statistics used in an ABC approach that requires a sophisticated forward model relating known isotope abundances to detected peak area [28,29]. The nuclear data that enter any analysis approach (ABC or other methods) include gamma emission energies, branching ratios, and half-lives. The branching ratios and half-lives determine the relative intensities of each peak for a given Pu isotopic fraction. Reference [29] indicates that uncertainties in emission energies are not as important in microcalorimetry as they are in lower resolution gamma spectroscopy such as that obtained in high-purity Germanium detectors, where spectral deconvolution is more challenging. ABC is compelling in spectrometry because ABC

requires user-chosen summary statistics such as estimated peak areas, ABC can easily accommodate uncertainty in nuclear data, ABC can provide an estimate of the posterior distribution of each unknown parameter, including the unknown isotopic abundances. However, ABC requires a good-quality forward model linking the summary statistics to the isotopic abundances as well as to fundamental nuclear data that has recognized uncertainties.

Only when there is good agreement between bottom-up and top-down UQ can the potential to improve an NDA method be fully understood. Many NDA methods require calibration, so one type of bottom-up UQ involves calibration data. Although calibration might appear to be a simple application of regression, [9,14] illustrate that simulation is needed for effective bottom-up UQ in NDA because sample sizes are small, a ratio of random variables in the calibration analysis is used, and there are non-negligible error variances in predictor and response. In addition, calibration data should include item-specific effects that will be present in testing data. As illustrated for the EMP, ABC is a good tool for bottom-up UQ. Once improved bottom-up UQ is implemented, any remaining disagreement between bottom-up and top-down UQ could indicate, for example, that there are missing sources of uncertainty in bottom-up UQ [20], that the data and/or error model are not what are assumed, or that correlations among inputs in the measurand equation (Eq. (2)) are not adequately estimated.

ABC is also effective for top-down UQ, for example, in paired (*O*,*I*) data. The advantages of a modern Bayesian approach applied to paired (*O*,*I*) data include the facts that one can: (1) accommodate any prior and any likelihood; (2) enforce any type of constraint, such as ratios of variances, with appropriate choice of prior, and (3) assess whether an implementation is well calibrated; for example, simulation can assess what fraction of 95% posterior probability intervals actually contain the true parameter such as σ^2_{Rd}. Disadvantages of a Bayesian approach include: (1) bias has to be assessed by sensitivity studies that vary the true and assumed likelihood and/or prior, and (2) numerical approaches such as Markov Chain Monte Carlo are easy to implement, but the user must perform convergence diagnostics to check whether one is really sampling from the correct posterior. ABC does not avoid such convergence issues, but the illustrated simulation strategy allows one to assess whether the chosen summary statistics, the distance measure, and the acceptance threshold lead to a well-calibrated approach.

7. References

[1] Burr, T., Hamada, M.S., Revisiting statistical aspects of nuclear material accounting Science and Technology of Nuclear Installations, Vol. 2013, Article ID 961360, 15 pages, doi:10.1155/2013/961360, 2013.

[2] Bonner, E., Burr, T., Guzzardo, T., Norman, C., Zhao, K., Beddingfield, D., Geist, W., Laughter, M., Lee, T., Improving the effectiveness of safeguards through comprehensive uncertainty quantification, Journal of Nuclear Materials Management, 44(2), 53-61, 2016.

[3] Burr, T., Krieger, T., Norman, C., Zhao, K., The impact of metrology study sample size on verification samples calculations in IAEA safeguards, European Journal Nuclear Science and Technology, 2, 36, 2016.

[4] Speed, T., Culpin, D., The role of statistics in nuclear materials accounting: issues and problems, Journal of the Royal Statistical Society B, 149(4), 281–313, 1986.

[5] Goldman, A., Picard, R., Shipley, J., Statistical methods for nuclear materials safeguards: an overview, Technometrics, 24(4), 267–275, 1982.

[6] Picard, R., Sequential analysis of materials balances, Nuclear Materials Management, 15(2), 38–42, 1987.

[7] Burr, T., Hamada, M.S., Bayesian updating of material balances covariance matrices using training data, International Journal of Prognostics and Health Monitoring, 5 (1) 006, pages: 13, 2014 .

[8] Walsh, S., Burr, T., Martin, K., The IAEA error approach to variance estimation for use in material balance evaluation and the international target values, and comparison to metrological definitions of precision, Journal of Nuclear Materials Management, 45(2):4-14, 2017.

[9] Burr, T., Croft, S., Jarman. K., Nicholson, A., Norman, C., Walsh, S., Improved uncertainty quantification in nondestructive assay for nonproliferation, Chemometrics, 159, 164-173, 2016.

[10] JCGM 104:2009, Evaluation of measurement data- an introduction to the “Guide to the Expression of Uncertainty in Measurement,” 2009.

[11] Fearnhead P., Prangle, D., Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation, Journal of the Royal Statistical Society B, 74(3), 419–474, 2012.

[12] Joyce, P., Marjoram, P., Approximately sufficient statistics and Bayesian computation, Statistical Applications in Genetics and Molecular Biology, 7(1), article 26, 2008.

[13] Burr, T., Skurikhin, A., Selecting summary statistics in approximate Bayesian computation for calibrating stochastic models, BioMedical Research International, Vol. 2013, Article ID 210646, 10 pages, 2013. doi:10.1155/2013/210646.

[14] Burr, T.,Croft, S., Krieger, T., Martin, K., Norman, C. Walsh, S., Uncertainty quantification for radiation measurements: bottom-up error variance estimation using calibration information, Applied Radiation and Isotopes, 108,49-57, 2015.

[15] ASTM C1514, Standard test method for measurement of ^{235}U fraction using the enrichment meter principle, 2008.

[16] Walton, R., Reilly, T., Parker, J., Menzel, J., Marshall, E., Fields, W., Measurements of UF6 cylinders with portable instruments, Journal of Nuclear Technology 21 (2), 133-148, 174.

[17] Miller, R., Beyond ANOVA: basics of applied statistics, Chapman & Hall, 1998.

[18] ISO 21748:2010 Guidance for the use of repeatability, reproduciblity, and trueness estimates in measurement uncertianty estimation.

[19] Zhao, K., Penkin, P., Norman, C., Balsely, S., Mayer, K., Peerani, P., Pietri, P., Tapodi, S., Tsutaki, Y., Boella, M., Renha G., Kuhn, E., STR-368 International target values 2010 for measurement uncertainties in safeguarding nuclear materials, IAEA, Vienna, 2010, www.inmm.org.

[20] Thompson, M., Ellison, S., Dark uncertainty, Accreditation and Quality Assurance 16, 483–487, 2011.

[21] Bich, W., Revision of the guide to the expression of uncertainty in measurement. Why and how, Metrologia 51, S155-S158, 2014.

[22] Elster, C., Bayesian uncertainty analysis compared to the application of the GUM and its supplements, Metrologia 51, S159-S166, 2014.

[23] Carlin, B., John, B., Stern, H., Rubin, D., Bayesian Data Analysis (1st ed), Chapman and Hall, 1995.

[24] Blum, M., Nunes, M., Prangle, D. Sisson, S., A comparative review of dimension reduction methods in approximate Bayesian computation. Statistical Science 28(2), 189–208, 2013.

[25] Nunes, M. Prangle, D., abctools: an R package for tuning approximate Bayesian computation analyses. The R Journal 7(2), 189–205, 2016.

[26] R: A language and environment for statistical computing, R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2008.

[27] Burr, T., Croft, S., Dale, D., Favalli, A., Weaver, B., Williams, B., Emerging applications of bottom-up uncertainty quantification in nondestructive assay, ESARDA Bulletin 53, 54-61, 2015.

[28] Burr, T., Croft, S., Hoover, A., Rabin, M., Exploring the impact of nuclear data uncertainties in microcalorimetry using ABC, Nuclear Data Sheets 123, 140-145, 2015.

[29] Hoover, A., Winkler, R., Rabin, M., Vo, D., Ullom, J., Bennett, D.,Doriese, W., Fowler, J., Horansky, R., Schmidt, D., Vale, L., Schaer, K., Determination of plutonium isotopic content by microcalorimeter gamma-ray spectroscopy, IEEE Transactions on Nuclear Science 60(2), 681-688, 2013.