An Overview of Statistical Methods used in Nuclear Safeguards

Tom Burr[1], Elisa Bonner[2], Sarah Michalak[1], Claude Norman[3]
[1]Los Alamos National Laboratory, Los Alamos NM 87545
[2]Colorado State University, Fort Collins CO 80523
[3]International Atomic Energy Agency, Vienna Austria A1220

Abstract
Nuclear safeguards at declared facilities aim to verify that nuclear material (NM) is used exclusively for peaceful purposes. To ensure that States honor safeguards obligations, measurements of NM inventories and flows are needed. Statistical analyses used to support conclusions require uncertainty quantification (UQ), usually by estimating the relative standard deviation (RSD) in random and systematic errors of each measurement method. A model is used for the normal (no facility misuse) data, and for the effects of NM misuse so that detection probabilities (DPs) for misuse scenarios can be estimated. This paper reviews statistical methods for UQ of measurements, for constructing tolerance intervals for setting pass/fail criteria for monitored data streams, and for estimating DPs for specified NM misuse scenarios at declared facilities. UQ for measurements is done both empirically using data collected for metrology studies and from applying error variance propagation to all steps in the assay (physics-based). Approximate Bayesian computation (ABC) is used for both the empirical and physics-based UQ. The estimated measurement error RSDs are then used to estimate the SD of the NM mass balances that are analyzed sequentially over time.

Key Words: approximate Bayesian computation, detection probability, uncertainty quantification, standard deviation

1. Introduction and Background

The nuclear industry must account for NMs such as items that contain plutonium-239, uranium-233, or uranium enriched in the 235 or 233 isotopes. Such materials are expensive, plus play a role in weapons production, so there are national and international security considerations. Domestic safeguards (note that "safeguards" can be singular or plural) protects NM and monitors for possible NM loss. International safeguards monitors for possible NM loss or diversion. Prominence of safeguards issues has generated attention by the general press such as the Wall Street Journal, New York Times, and Science; also, many papers on statistical methods and applications to safeguards problems have appeared in various journals [1-25].

Domestic safeguards has two main components: physical security to monitor and control access to NM, and accounting procedures to monitor the flow and location of NM quantities. Analogously, the banking industry also uses physical security (guards, cameras, locked vaults, etc.) together with accounting procedures to monitor money flow. However, monitoring NM is more complicated due to the multitude of physical forms and chemical compositions of NM in bulk quantities, resulting in measurement errors. In international safeguards, inspectors recognize that the facility could divert NM and alter

the accounting records to try to mask the diversion. In contrast, domestic safeguards does not include the possibility that the facility might alter the accounting data to mask NM diversion. And, in contrast to domestic safeguards, international safeguards makes no attempt to control the NM.

NM control and accountability have been a concern since the beginning of the atomic era. Establishment of the International Atomic Energy Agency (IAEA) in 1957 represented a landmark in international safeguards cooperation. The IAEA, by international agreement, provides verification to monitor and ensure the peaceful use of nuclear materials and technology around the world. An account of the evolution of international safeguards is given by Willrich [24]. A more recent account [21] describes evolution of the "state-level" concept that resulted from advancement of safeguards concepts after discovery of an undeclared nuclear facility in the early 1990s (which lead to expanded country-wide monitoring to monitor for undeclared NM activities).

In nuclear material accountancy (NMA) at declared facilities, the material balance (MB) is defined as $MB = I_{begin} + T_{in} - T_{out} - I_{end}$, where T_{in} is transfers in, T_{out} is transfers out, I_{begin} is beginning inventory, and I_{end} is ending inventory. The measurement error standard deviation of the MB is denoted σ_{MB}. The key quantities in NMA are the MB and σ_{MB}. If the MB at a given time ("balance period") exceeds $k\ \sigma_{MB}$ with k in the 2-to-3 range, then the NMA system "alarms." Considerable effort is aimed at assessing measurement uncertainties ("uncertainty quantification," UQ) to estimate σ_{MB} [1-4,6-9,13-23]. Choosing k in the 2-to-3 range for a low false alarm probability is based on an appeal to a central limit theorem effect arising from combining many measurements to justify assuming the measured MB is approximately Gaussian distributed around the true MB [1-5,14,18,20]. The IAEA must "trust, but verify," which means that a sample of the operator's declared and measured items will be re-measured by inspectors, often using portable non-destructive assay (NDA) methods based on detecting and interpreting gamma and/or neutron emissions from the item [1,7,8].

This paper reviews statistical methods for UQ of measurements (Section 2), for constructing tolerance intervals for setting pass/fail criteria for differences between operator declarations and corresponding inspector measurements (Section 3), and for estimating DPs for sequences of MBs at declared facilities (Section 4).

2. UQ

2.1 Measurement Error Models

Statistical analyses used to support conclusions require UQ, usually by estimating the relative standard deviation (RSD) in random and systematic errors associated with each measurement method [1-9]. UQ for measurements is done both empirically using data collected for metrology studies and from applying error variance propagation to all steps in the assay (physics based). Approximate Bayesian computation (ABC) can be used effectively for both the empirical and physics-based UQ (see below).

A defensible measurement error model for operator and inspector data must account for variation within and between groups, where a group is, for example, a calibration period, inspection period, item, or laboratory. In this paper, a group is an inspection period. A typical model that assumes additive errors for the inspector's measurements (I_{jk}) (and similarly for the operator O_{jk}) is

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

where for item k (from 1 to n) in group j (a group is an inspection period in this paper, from 1 to g), I_{jk} is the inspector's measured value of the true but unknown value μ_{jk}, $R_{jk} \sim N(0, \sigma_{RI}^2)$ is a random error in the inspector (I denotes inspector) measurement,

and $S_{Ij} \sim N(0, \sigma_{SI}^2)$ is a short-term systematic error that arises due to metrology changes, the most important of which is recalibration between inspection periods [1-3,14,26,27]. For a fixed value of μ_{jk}, the total variance of the inspector measurement is $\sigma_I^2 = {}_{SI}^2 + \sigma_{RI}^2$, assuming that random and systematic errors are independent. This paper combines all inspector measurements into one "inspector" group, even if there are two or more inspectors.

The measurement error model in Eq. (1) sets the stage for applying ANOVA with random effects [2,16-19]. Neither R_{Ijk} nor S_{Ij} are observable. If the errors tend to scale with the true value, then a typical model for multiplicative errors (with relative standard deviations (RSD) δ_S and δ_R) is

$$I_{jk} = \mu_{jk}(1 + \tilde{S}_{Ij} + \tilde{R}_{Ijk}) \tag{1}$$

where $\tilde{S}_{Ij} \sim N(0, \delta_{SI}^2)$, $\tilde{R}_{Ijk} \sim N(0, \delta_{RI}^2)$. As explained below, for a technical reason, the data model in Eq. (1) is slightly modified to use truncated normal distributions instead of normal distributions in the IAEA application. Let $\delta_R^2 = \delta_{RO}^2 + \delta_{RI}^2$ and $\delta_S^2 = \delta_{SO}^2 + \delta_{SI}^2$, where again "O" denotes operator and "I" denotes inspector. For a fixed value of μ_{jk}, the total variance of the inspector measurement is $\sigma_I^2 = \sigma_{SI}^2 + \sigma_{RI}^2 = \mu_{jk}^2(\delta_{SI}^2 + \delta_{RI}^2)$. Let $O_{jk} = \mu_{jk}(1 + \tilde{S}_{Oj} + \tilde{R}_{Oj})$ and $I_{jk} = \mu_{jk}(1 + \tilde{S}_{Ij} + \tilde{R}_{Ij})$. Subsequently, the assumed model for the relative difference between operator and inspector is

$$\frac{\mu_{jk}(1 + \tilde{S}_{Oj} + \tilde{R}_{Ojk}) - \mu_{jk}(1 + \tilde{S}_{Ij} + \tilde{R}_{Ijk})}{\mu_{jk}} = \tilde{S}_j + \tilde{R}_{jk} \tag{2}$$

for the operator's declared value of item k from group j, $\tilde{R}_{jk} = \tilde{R}_{Ojk} - \tilde{R}_{Ijk} \sim N(0, \delta_R^2)$. is the net random error and $\tilde{S}_j = \tilde{S}_{Oj} - \tilde{S}_{Ij} \sim N(0, \delta_S^2)$ is the net short-term systematic error. In practice, while assuming no data falsification by the operator, Eq. (2) can be calculated using the relative differences, $d_{jk} = \frac{O_{jk} - I_{jk}}{O_{jk}}$ where o_{jk} is used in the denominator to estimateμ_{jk}, because typically $\delta_{SO}^2 + \delta_{RO}^2 \ll \delta_{SI}^2 + \delta_{RI}^2$, with $\delta_{TO} = \sqrt{\delta_{SO}^2 + \delta_{RO}^2}$ always being very small, 0.02 or less, in the IAEA application. The technical issue mentioned above is that a ratio of normal random variables has infinite variance [3, 21]. To define a ratio that has finite variance, a truncated normal can be used as the data model in Eq. (2) for o_{jk} in $d_{jk} = 1 - \frac{I_{jk}}{O_{jk}}$, which is equal in distribution to $1 - \frac{\mu_{jk}(1+\delta_{TI}z_1)}{\mu_{jk}(1+\delta_{TO}z_2)}$, which involves a ratio $R = \frac{(1+\delta_{TI}z_1)}{(1+\delta_{TO}z_2)}$ of the independent normal random variables $z_1$and z_2 (for the case of one measurement per group; multiple measurements per group is treated similarly). Because δ_{TO} is so small, usually 0.02 or less, such truncation (at approximately 50 times $\delta_{TO} = 0.02$ for example) has no noticeable effect, but ensures that the ratio *R* has finite moments [4].

2.3 UQ for NDA

NDA uses calibration and/or modelling to infer 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 [26]

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 [27]. 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 [28] 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.

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

$$Y = f(X_1, X_2, \ldots, X_N) \qquad (4)$$

for measurand Y and inputs $X_1, X_2, \ldots, X_N$. The GUM applies the delta method to Eq. (4) 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. (4) can be well approximated by a first-order Taylor series expansion around the mean values of each X_i, 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 X_i can be accommodated). If the first-order Taylor series is not sufficiently adequate, the GUM recommends Monte Carlo simulation. Note that Eq. (4) implies that Y is random, so the GUM implicitly adopts a Bayesian viewpoint (Section 4) without explicitly stating a prior distribution for Y [7,8,10].

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 [7,8]. Toward that goal, several U.S. national laboratories and the standards (ASTM) committees are 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 [29].

2.3 ABC for top-down UQ

Bayesian ANOVA such as could be applied to data generated from Eq. (1) has been studied [30]; 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. 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. So 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 [31]. For numerical methods such as Markov Chain Monte Carlo [31], the user specifies a prior distribution for δ_{RI}^2 and δ_{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.

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 summarized as follows:

ABC Inference
For i in 1, 2, …, N, do these 3 steps:
(1)Sample θ from the prior, $\theta \sim f_{\text{prior}}(\theta)$.
(2) Simulate data x' from the model $x' \sim P(x \mid \theta)$
(3) Denote the real data as x. If distance
$d(S(x'), S(x)) \leq \varepsilon$, accept θ as an observation from $\text{f}_{\text{posterior}}(\theta \mid x)$.

Experience with ABC suggests that the ABC approximation to $\text{f}_{\text{posterior}}(\theta \mid x)$ improves if step (3) is modified to include a weighting factor, so that trial values of θ simulated from $f_{prior}(\theta)$ that lead to very small distance $d(S(x'), S(x))$ are more heavily weighted in the estimated posterior [32-36]. In step (2), the model can be analytical or, for example, a forward transport model.

In ABC, the model has input parameters θ and outputs data $x(\theta)$ and there is corresponding real data x_{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 $\theta|\ \text{y}_{\text{obs}}$ if the distance $d(\text{x}_{\text{obs}}, \text{x}(\theta))$ between x_{obs} and $\text{x}(\theta)$ is reasonably small. Alternatively, for most applications, it is necessary to reduce the dimension of x_{obs} to a small set of summary statistics S and accept trial values of θ if $d(S(\text{xobs}), \text{S}(\text{x}(\theta)) < \varepsilon$, where ε is a user-chosen threshold. Here, for example, $x_{obs} = d_{rel} = \frac{O-I}{O}$ data in each inspection group, and $S(\text{x}_{\text{obs}})$ includes within and between groups sums of squares Specifically, the ANOVA-based estimator of δ_{RI}^2 in Eq. (1) is $\hat{\delta}_R^2 = \frac{1}{ng-g}\sum_{j=1}^{g}\sum_{k=1}^{n}(d_{jk} - \bar{d}_j)^2$, and the usual estimate of δ_S^2 is $\hat{\delta}_S^2 = \frac{\sum_{j=1}^{g}(\bar{d}_j - \bar{\bar{d}})^2}{(g-1)} - \frac{\hat{\delta}_R^2}{n}$, where $\bar{\bar{d}} = \frac{1}{ng}\sum_{j=1}^{g}\sum_{k=1}^{n} d_{jk}$ is the overall unweighted average. 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 $d(S(x_{\text{obs}}), \text{S}(\text{x}(\theta)) < \varepsilon$ 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 $d(S(\text{xobs}), \text{S}(\text{x}(\theta)))$ (abctools in R [36,37]).

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 and from their prior distribution $p_{\text{prior}}(\theta)$ where $\theta = (\delta_R^2, \delta_R^2)$; (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 $d(S(\text{xobs}), \text{S}(\text{x}(\theta)) < \varepsilon$. If desired, aiming to improve the approximation to the posterior, adjust the accepted θ values on the basis of the actual $d(S(\text{xobs}), \text{S}(\text{x}(\theta))$ 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 [32] 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 [34] illustrated a method to evaluate whether a candidate set of summary statistics leads to a well-calibrated posterior, in the same sense used here; 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.

2.4 ABC for bottom-up UQ

This section describes ABC for the enrichment meter principle (EMP), that is based on the count rate of a 185.7keV gamma-ray. The mass of ^{235}U in an item can be estimated by using a measured net weight of uranium U in the item and a 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 [7,8,28,38]. 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 must be homogeneous with respect to both the ^{235}U enrichment and chemical composition, and that the container attenuation of gamma-rays is that 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 [7,8,28,38] 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 [28] (the RSD tends to be larger for lower 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 energy channels somewhere below and above the 185.7 keV energy but outside the peak area to estimate background (two-region EMP method), expressed as

$$y = \beta N + R_Y \qquad (5),$$

where Y is the enrichment, N is the peak count rate near 185.7keV, R_Y is random error. Figure 1 is an example low-resolution (NaI detector) gamma spectrum near the 185.6keV. The gross count and the two background ROI counts can be combined into one net count, resulting in one predictor as in Eq. (5), which can be fit using least squares

regression. 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 [39]. Alternatively, both peak and background counts can be used as predictors [7,8,28,39]. 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 of reverse calibration (Eq. (5) 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 generate the estimate $\hat{\beta}_1$ of the model parameter β_1. The variance of β is not necessarily well-approximated by the usual least squares expression because of errors in *N* [7,8,39,40]. Therefore, [14,27] suggest that the root mean squared error (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 $E\left(\left(\hat{Y} - Y_{true}\right)^2\right) = E\left(\left(\hat{Y} - E(\hat{Y})\right)^2\right) + ((E\left(\hat{Y}\right) - Y_{true})^2 =$ variance + bias^2.

One can express the calibration Eq. (5) as in Eq. (4), where X_1 is $\hat{\beta}_1$ and X_2 is *N*, with var $(\hat{\beta}_1)$ estimated by simulation, so GUM's Eq. (4) 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. (4) is not actually designed to be applied to calibration applications, regardless of whether there are errors in the predictors.

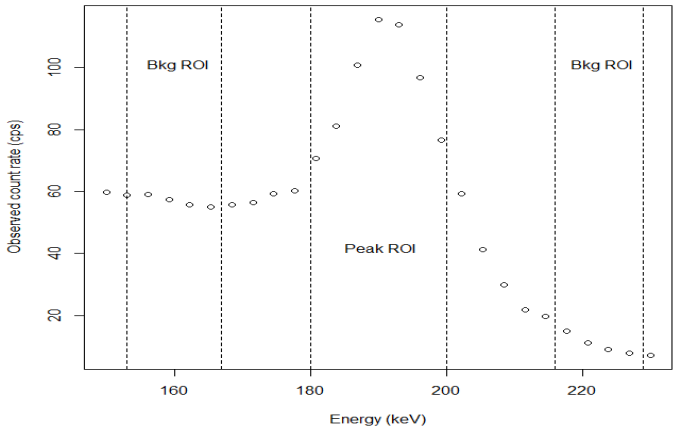

Figure 1. Example low-resolution (NaI detector) gamma spectrum near the 185.6keV peak with two background regions (one region below the 185.7 keV peak and one region above the 185.7 keV peak).

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

$$\frac{CR}{M} = g(X_1, X_2, \ldots, X_N) \quad (6),$$

where CR is the item's neutron or gamma count rate, M is the item SNM mass, g is a known function, and $X_1, X_2, \ldots, X_N$ are N auxiliary predictor variables such as item density, source SNM heterogeneity, and container thickness, which will generally be estimated or measured with error and so are regarded as random variables . To map Eq. (6), to GUM's Eq. (2), write

$$M = \frac{CR}{g(X_1, X_2, \ldots, X_N)}) = h(X_1, X_2, \ldots, X_M) \quad (7),$$

where the measured CR is now among the $M = N+1$ inputs. Note that Eq. (7) is the same as Eq. (2), but some of the X_i account for item-specific departures from reference items used for calibration. More specifically, Eq. (5) can be re-expressed as

$$y = \beta_{item} N + R_Y \quad (8),$$

where the calibration "constant" denoted β_{item} varies across items. Equation (8) is a random-coefficient regression equation, and real and/or simulated data generated from Eq. (8) can be used to estimate the average value of β_{item}. Eq. (8) 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. (8). And, one way to model the effect of imperfect adjustment for each item's container thickness is to include another random error in the net count rates rather than to modifyβ_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 $e^{-\mu\rho t}$). Note that errors in N have the same impact as errors in β_{item}because the term $\beta_{item}N$ appears in Eq. (8).

2.4 ABC applied to the EMP for bottom-up and top-down UQ

ABC applied to the EMP for bottom-up UQ can be implemented in the following seven steps. (1) Estimate the average regression coefficient in Eq. (8) using available real calibration data, typically consisting of approximately 3 to 5 (Y,N) pairs. The example real calibration data used here are Y = 0.355, 0.80, 2.175, .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 to generate synthetic sets of 5 paired (Y,N) values, with run i producing the estimate $\hat{\beta}_{1,i}$. This example generated the β_{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 a wide 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 true mean net count rate that corresponds to a $\mu_Y$value sampled from its prior distribution. This example used an RSD in Y of 0.1% and in of 5%. (5) Specify a count time (this example used 600 seconds) t, simulate $B \sim \text{Pois}(\mu_B t)$, and $G \sim \text{Pois}(\mu_G t)$, and compute a net count rate (this example assumes 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*. The net count rate *N* is an effective summary statistic. (7) For each simulated test case, simulate a value of 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 generated in step (6) that correspond to $= |N_{test} - N_i| \leq \varepsilon$ as observations from the posterior $\mu_Y|N$ (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 $\mu_Y$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.

To assess ABC performance, the two criteria mentioned can 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 be in good agreement with the actual coverage. The analysed data 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 the d_{rel}=$\frac{O-I}{O}$ which includes both within and between group standard deviations) from top-down data (also using ABC for the top-down ANOVA) is 0.11, which is very close to that predicted from the bottom-up ABC-based (0.12 as explained in the next paragraph) posterior standard deviations for *O* and *I*. Recall that the usual ANOVA-based estimator of δ_R^2(using the multiplicative form of Eq. (1) for both operator and inspector) is $\hat{\delta}_R^2 = \frac{1}{ng-g}\sum_{j=1}^{g}\sum_{k=1}^{n}(d_{jk} - \bar{d}_j)^2$,and the usual estimate of δ_S^2 is $\hat{\delta}_S^2 = \frac{1}{ng-g}\frac{\sum_{j=1}^{g}(\bar{d}_j - \bar{\bar{d}})^2}{(g-1)} - \frac{\hat{\delta}_R^2}{n}$. The quantities, $\hat{\delta}_R^2$and $\hat{\delta}_S^2$are therefore good summary statistics for ABC, and were used to implement ABC for top-down UQ.

The 0.12 bottom-up prediction for the RSD δ_d of d (computed here as the SD of d_{rel}=$\frac{O-I}{O}$) is illustrated by plotting the posterior for *O* for a particular *N* value in Fig. 2, 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}=$\frac{O-I}{O}$is given by$\sqrt{0.08^2 + 0.08^2} = 0.11$. The 0.12 top-down estimate of the RSD of δ_d (see Fig. 3, which is based on g = 3 groups and n = 10 paired measurements per group) is the RSD of the ABC-based posterior distribution for $\delta_d$from top-down UQ. The 0.12 estimate has a SD of 0.03, and an approximate 95% probability interval for $\delta_d$is 0.07 to 0.21. The estimated posterior for $\delta_d$has approximately the same mean and SD regardless of whether the data are generated from Eq. (1) or from the bottom-up-based measurement method. Figure 4 combines Figures 2 and 3 to show that the bottom-up and top-down posteriors for $\delta_d$are in good agreement.

One advantage of having a probability interval for both the bottom-up and top-down estimate of δ_d is that one can assess whether differences between the top-down and bottom-up estimates of $\delta_d$are significant. In this example, bottom-up UQ using ABC agrees well with corresponding top-down UQ using ABC that used simulated *O* and *I* values as in Fig. 1. 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 unmodeled effects ("dark uncertainty" [26]), 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 (Fig. 4), there is no evidence of dark uncertainty (and there should not be, because no dark uncertainty was simulated).

A normal distribution is not always a good approximation for the actual distribution of $\frac{O-I}{O}$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 lamba 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.

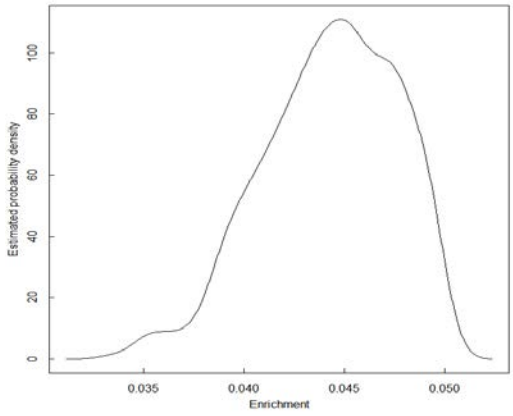

Figure 2. The bottom-up ABC-based estimate of the posterior δ_{TI}(or δ_{TO}).

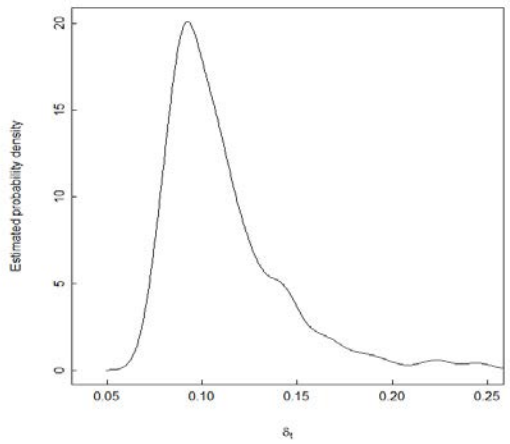

Figure 3. The top-down ABC-based estimate of the posterior for δ_d with mean 0.12 and RSD of 0.03.

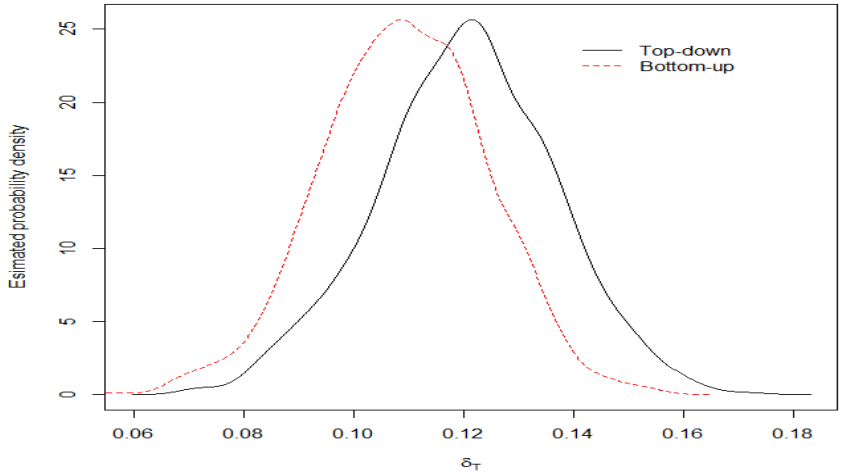

Figure 4. The bottom-up and top-down ABC-based posterior estimates of δ_d.

3. Tolerance Intervals

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 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, typically defined as $D_{abs} = \frac{N}{n}\sum_{j=1}^{n}(O_j - I_j)$, or as $D_{rel} = \frac{N}{n}\sum_{j=1}^{n}\frac{O_j - I_j}{O_j}$, 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. 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].

Many applications involve testing for a shift in the mean of a probability distribution for a random quantity, with a false alarm probability (FAP) and a failure-to-detect-the-shift probability. The IAEA verification data that are collected to monitor for possible nuclear material diversion is the example used here [1,4]. Such verification data from IAEA inspections often consist of paired data (usually operators' declarations and inspectors' verification results) that are analyzed to detect the significant differences. Any significant difference could arise due to a problem with the operator and/or inspector measurement systems, or due to the operator falsifying the data in an attempt to mask diversion; such a falsification would cause a mean shift between the operator declaration and the corresponding inspector measurement. Paired data from past inspections are used to estimate the alarm limits, and each inspection period is regarded as a group within which the measurements are assumed to have the same systematic error. due to calibration and other effects [1–3]. The corresponding FAP depends on the assumed measurement error model and its random (within-group) and systematic (between-group) error variances, which are estimated while using data from previous inspections [3].

Reference [4] reviewed parametric, semi-parametric, and non-parametric options for setting alarm thresholds in such grouped data. If both the within-group and between-group measurement errors have approximately normal distributions, then a parametric option that involves tolerance intervals [4-11] for one-way ANOVA can be used. If either or both error distributions are not close to normal, then a semi-parametric method that is based on a Dirichlet process mixture [4] can be applied. A non-parametric method [4] could be used if there is enough data.

3.1 Parametric approach

TI construction methods for one-way ANOVA continue to be improved. In some applications, all four error variance components in d_{jk} from Section 2 must be estimated [1–3,17–19], but in this application, only the aggregate variances δ_S^2 and δ_R^2 need to be estimated. This paper's focus is on the total relative variance, $\delta_T^2 = \delta_S^2 + \delta_R^2$, because $d_{jk} \sim N(0, \delta_T^2$ (approximately, due to using $d_{jk} = \frac{o_{jk} - i_{jk}}{o_{jk}}$ rather than $d_{jk} = \frac{o_{jk} - i_{jk}}{\mu_{jk}}$). The estimated variances $\hat{\delta}_R^2$ and $\hat{\delta}_S^2$ are used to compute $\hat{\delta}_T^2 = \hat{\delta}_S^2 + \hat{\delta}_R^2$, so that $\hat{\delta}_T$ can be used to set an alarm threshold for future d_{rel} values. Specifically, in future values of the operator-inspector difference statistic $d = \frac{o-i}{o}$, if $|d| > k\hat{\delta}_T$ (in two-sided testing), then the *i*-th item selected for verification leads to an alarm, where $\hat{\delta}_T = \sqrt{\hat{\delta}_S^2 + \hat{\delta}_R^2}$ (with $\hat{\delta}_T$ the total RSD, δ_S the between-period short-term systematic error RSD, and δ_R the within-period reproducibility) and k = 3 is a common current choice that corresponds to a small α of approximately 0.003 if $\hat{\delta}_T = \delta_T$. Therefore, the focus in [4,25] is a one-way random effects ANOVA

[30]. Regarding jargon, note that the short-term systematic errors are fixed within an inspection period, but they are random across periods, so this is called a random effects ANOVA model [30]. Due to the estimation error in $\hat{\delta}_T$, the actual FAP can be considerably larger than 0.05, as shown in [10,25].
The usual ANOVA decomposition is

$$\sum_{j=1}^{g}\sum_{k=1}^{n}(d_{jk}-\bar{\bar{d}})^2=\sum_{j=1}^{g}\sum_{k=1}^{n}(d_{jk}-\bar{d}_j)^2+n\sum_{j=1}^{g}(\bar{d}_{j.}-\bar{\bar{d}})^2=\text{SSW}+\text{SSB}=S_W^2+S_B^2 \quad (9)$$

where $d_{jk}=o_{jk}-i_{jk}$ for additive models and $d_{jk}=\frac{o_{jk}-i_{jk}}{o_{jk}}$ for multiplicative models as assumed from now on (to avoid cluttering the notation, the "rel" subscript in $d_{rel}=\frac{O-I}{O}$ is omitted). In Eq. (9), SSW is the within group sum of squares, and SSB is the between group sum of squares. For simplicity, Eq. (9) assumes that each group has the same number of measurements n. As usual in one-way (one grouping variable) random effects ANOVA, $E\left(\sum_{j=1}^{g}\sum_{k=1}^{n}(d_{jk}-\bar{d}_j)^2\right)=(ng-1)\delta_R^2$ and $E\left(\frac{n}{g-1}\sum_{j=1}^{g}(\bar{d}_{j.}-\bar{\bar{d}})^2\right)=n\delta_S^2+\delta_R^2$ from which it follows that reasonable estimators are $\hat{\delta}_R^2=\frac{S_W^2}{g(n-1)}$ and $\hat{\delta}_S^2=\frac{1}{n}(\frac{S_B^2}{g-1}-\frac{S_W^2}{g(n-1)})$.

For data sets in which d_{jk} appears to have approximately a normal distribution, key properties (such as the variances) of the estimators $\hat{\delta}_R^2$ and $\hat{\delta}_S^2$ are approximately known [30]. However, biased estimators can have smaller mean squared error (MSE), so other estimators should be considered. Additionally, again assuming normally distributed *R* and *S* values, an exact confidence interval (CI) can be constructed for δ_R^2 using the $\chi^2_{df=g(n-1)}$ distribution, but there are only approximate methods to construct CIs for δ_R^2, because the distribution of $\hat{\delta}_S^2$ is a difference of two independent χ^2 random variables. Kraemer [41] proposes a modified CI construction method for δ_S^2 and investigates impacts of non-normality.

Many readers are probably more familiar with confidence intervals (CIs) than tolerance intervals (TIs). A CI is defined as an interval that, on average, includes a model parameter, such as a population mean with a stated confidence, often 95%. A TI is very similar to a CI, but it is defined as in interval that bounds a percentage of the population with a stated confidence, often bounding 95% of the population with confidence 99%. In the IAEA application, an alarm threshold is used that is assumed to correspond to a small false alarm probability (FAP), such as 5%, so the TI-based threshold bounds the lower (one-sided testing) or middle (two-sided testing) 95% of the population. Therefore, TIs are needed to control the FAP with high confidence (such as 99%) to be 5% or less.

Historical differences, such as $d_1,\ d_2,\ldots,\ d_{ng}$ are often used to estimate an alarm threshold for future measurements that has a small nominal α, such as $\alpha=0.05$. Accordingly, instead of requiring a CI for the true value μ the need is to estimate a threshold denoted $T_{0.95}$, which is the 0.95 quantile of the probability

distribution of d corresponds to $\alpha = 0.05$ in one-sided testing. In contrast to a CI, a TI is an interval that bounds a fraction of a probability distribution with a specified confidence (frequentist) or probability (Bayesian approach) [4,10,31], in this paper, for the model $X = \mu + S + R$, where $X = \frac{O-I}{O}$ as computed with paired (O,I) data.

It is helpful to first review inference in a simpler setting without data being grouped by an inspection period. Suppose that data $x_1, x_2, \ldots, x_n, x_{n+1}$ are collected from a distribution that is approximately normal with unknown mean μ and standard deviation σ, so $X \sim N(\mu, \sigma)$. Assume that x_{n+1} is test data and that $x_1, x_2, \ldots, x_n$ are the training data used to estimate μ and σ, while using the usual estimates $\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\hat{\sigma} = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$, respectively. When constructing intervals of the form $\bar{x} \pm k\hat{\sigma}$, the multiplier k can be chosen in order to have any user-desired confidence that the interval $\bar{x} \pm k\hat{\sigma}$ will include the true parameter μ Specifically, for the commonly-used t-based CI,
$k = t_{1-\frac{\alpha}{2}}(n-1)$ where $1-\alpha$ is the desired confidence and $t_{1-\frac{\alpha}{2}}(n-1)$ denotes the
$1-\frac{\alpha}{2}$ quantile of the t distribution with n-1 degrees of freedom. For example, if n = 10, 20, or 30, then $k = t_{1-\frac{\alpha}{2}}(n-1)$ = 2.26, 2.09, or 2.05, respectively. Note that the well-known t-based CI is appropriate for ungrouped data. Or, if σ is known, then $k = z_{1-\frac{\alpha}{2}}$, where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the normal distribution (the commonly-used z-based CI).

The previous paragraph adopted a frequentist viewpoint (μ and σ are unknown constants), so the intervals are referred to as CIs. In repeated applications of training on n observations of X, a fraction of approximately $1-\alpha$ of these CIs for μ will include the true value of μ (and similarly for σ). The Bayesian viewpoint regards μ and σ as random variables. A prior distribution is assumed for both μ and σ and the training data are used via Bayes theorem to update the prior to produce a posterior distribution [31-35,37], which is used to produce an interval that includes the true parameter with any user-desired probability (assuming that the data $x_1, x_2, , , x_n,$ are approximately normal and the prior distributions for μ and σ are appropriate). The Bayesian approach is subjective, unless there is some objective means to choose the prior probability [31]; however, in the context of this paper, there are simple options to validate that the Bayesian approach is calibrated.

Moving from inference about μ and σ [ref] introduced methods to use $x_1, x_2, \ldots, x_n$ to calculate a threshold $\hat{T}_{1-\alpha}$ such that $P(\hat{T}_{1-\alpha} \geq T_{1-\alpha})$, where $T_{1-\alpha}$ is the true $1-\alpha$ threshold of the distribution of x. Frequentist TIs and Bayesian prediction intervals were developed, which include a specified fraction of at least $1-\alpha$ of future data, with p being the specified confidence in the frequentist TI approach and p being the posterior probability in a Bayesian approach [10]. The

frequentist TI estimators that are presented have the form $\hat{T}_{0.95} = \hat{\mu} + k\hat{\sigma}$, where k is the coverage factor that depends on the sample size. In any Bayesian approach, probabilities are calculated with respect to the joint posterior distribution $f_{posterior}(\mu, \sigma)$ for given $x_1, x_2, \ldots, x_n$ [22]. In this context, in the Bayesian approach, μ and σ are random unknown parameters, so $P(\hat{T}_{0.95} \geq T_{0.95}) = P_{\mu,\sigma}(\hat{T}_{0.95} \geq T_{0.95})$ is computed with respect to μ and σ for given $x_1, x_2, \ldots, x_n$. The Bayesian approach that was used in the IAEA application generates observations μ and σ from the posterior probability, which can be used to compute the posterior means $\hat{\mu}$ and $\hat{\sigma}$ (the hat notation is somewhat non-standard in Bayesian literature, but it denotes the respective point estimate), which can then be used to numerically search for a suitable value of k to estimate $\hat{T}_{0.95} = \hat{\mu} + k\hat{\sigma}$. In the frequentist approach, $\hat{\mu}$ and $\hat{\sigma}$are random, while μ and σ are fixed unknowns, so $P(\hat{T}_{0.95} \geq T_{0.95}) = P_{X_1,X_1,\ldots,X_n}(\hat{T}_{0.95} \geq T_{0.95})$ is computed with respect to random samples of size n. A frequentist TI has an associated confidence, which is the long-run relative frequency that an interval such as $(0, \hat{T}_{0.95} = \hat{\mu} + k\hat{\sigma}$ will include a future observation X from the same distribution as the training data used to estimate $\hat{\mu}$ and $\hat{\sigma}$.

An exact expression for a TI is only available in the one-sided one-component Gaussian case [4,10]. However, good approximate expressions for many other cases are available [4]. Alternatively, TIs can be estimated well by using a simulation to approximate an alarm threshold that is designed to contain at least $1 - \alpha$ percent of future observations with a specified coverage probability p.

For the case $X \sim N(\mu, \sigma)$, in one-sided testing, [4] the exact upper limit for a 1-p TI upper bound is $U = \bar{x} + ks$ where $k = t_{1-\alpha}(\text{df} = n - 1, \text{ncp} = \lambda)/n$ where the noncentrality parameter $\lambda = z_p\sqrt{n}$, and $z_p$denotes the pth quantile (p = 0.99 here) of the standard normal (mean 0, variance 1). For example, with n = 10, 20, or 30, k = 3.74, 2.81, and 2.52, respectively. Note that these values of k are larger than the corresponding values in a $1 - \alpha$ CI for μ (2.26, 2.09, or 2.05, from above). results in order to illustrate the simulation approach.

In the ANOVA setting, the factor k required in $\hat{T}_{0.95} = \hat{\mu} + k\hat{\sigma}$ depends on the unknown ratio $\frac{\sigma_S}{\sigma_R}$, so approximations are needed. As an example, if one assumes $\hat{\delta}_T = \delta_T$, then choosing $= 1.65$ corresponds to $\alpha = 0.05$ (one-sided testing with the Gaussian approximation); however, if n = 10, paired measurements in each of g = 3 prior inspection periods are available, and $\frac{\sigma_S}{\sigma_R} = 1$, then choosing k = 1.65 leads to an actual FAP of 0.05 or less, with a probability of only 0.38. Therefore, one must choose a larger value of k than $k = 1.65$ in order to ensure a large probability p that $\alpha \leq 0.05$ [4,10]. Unlike the single-component Gaussian case above, the required value of k depend on the ratio $\frac{\sigma_S}{\sigma_R}$, which is unknown in practice, so approximate methods are used. If one desires a high probability p = 0.99 that the actual FAP is as small as the nominal FAP (0.05), then simulations in R [20] indicate that the required k values are 2.52, 2.94, or 4.23, for $\sigma_S = 0.25\sigma_R, 1\sigma_R, 4\sigma_R$ respectively. Not

surprisingly, the required k value increase as $\frac{\sigma_S}{\sigma_R}$ increases. Reference [4,40] have more detail on parametric TI construction for ANOVA.

3.2 Semiparametric approach

A semiparametric approach lies between the parametric and nonparametric extremes. For example, one semiparametric option is bspmma (which uses a dirichlet process prior) that can be compared to the parametric case, and for a two-sided interval expressed as $0 \pm k\hat{\sigma}_T$, requires k = 4.2 for the example data in Fig. 5 simulated from Eq. (1). The R code bspmma [42] can be used in one-way random effects ANOVA to estimate the posterior distribution of *S.* The acronym bspmma stands for Bayesian semiparametric models for meta-analysis [42], and the R code uses model selection that is based on the Radon–Nikodym derivative. The posterior distribution of R can be obtained by any of several common approaches. The $k = 4.2$ result for the $S + R$ distribution used a parametric bootstrap [4], in which sample i from the posterior distribution of $S + R$ used bspmma to generate the S value and from $N(0, \hat{\sigma}_{R_i})$ to generate the R value. The standard deviation $\hat{\sigma}_{R_i}$ is sampled from the posterior distribution of σ_{R_i} .

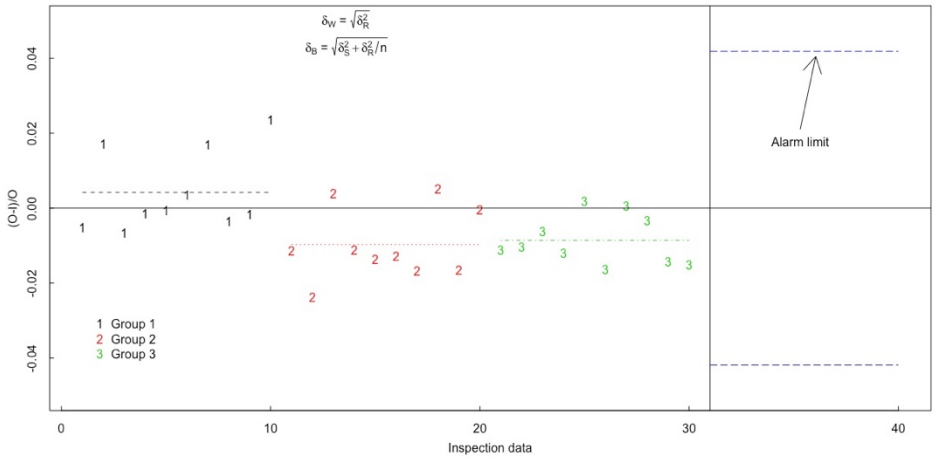

Figure 5. Simulated verification measurement data with $\delta_S = \delta_R = 0.01$. The relative difference $d = \frac{o-i}{o}$ is plotted for each of 10 paired (o,i) measurements in each of three groups, for a total of 30 relative differences. The mean relative difference within each group (inspection period) is indicated by a horizontal line through the respective group means of the paired relative differences. The 3 sets of 10 d values (multiplied by 100 for ease of reading) are {(−0.52, 1.72, −0.66, −0.15, −0.06, 0.35, 1.70, −0.35, −0.17, 2.35), (−1.14, −2.38,0.40, −1.12, −1.36, −1.28, −1.67, 0.52 −1.65, −0.03), (−1.12, −1.04, −0.63, −1.19, 0.19, −1.63, 0.07, −0.34, −1.43, −1.50)}.

The semi-parametric approach that was used in this context assumes that R has a normal distribution and S has an arbitrary (unknown) distribution. Interestingly, in this case, the covariance between MSB and MSW is 0, because $\text{cov}(MSB, MSW) =$

$\frac{\mu_{4,R}-3\sigma_R^4}{ng}$where $\mu_{4,R}$ is the fourth central moment, $\mu_{4,R} = E(R^4)$, and then because $\mu_{4,R} = 3\sigma_R^4$ for a normal distribution [30], cov(MSB, MSW)=0.

3.3 Nonparametric approach

The nonparametric approach requires one or more observations from each of the 130 groups (inspection periods), which are far too many groups and observations to be practical. The basis for the nonparametric result of n = 130 is an application of order statistics [4], and the expression for the required sample size n from the $S + R$ distribution in model 2 is $1 - np^{n-1} + (n-1)p^n$, where p is the desired probability. If the requirement is 95% probability that the tolerance interval that is based on that the interval $X_{(n)} - X_{(1)}$ = max – min contains at least 95% of future values, then n = 93. Raising the requirement from 95% probability to 99% probability increases the required n to n = 130, as given.

4. Detection Probabilities

NMA requires measuring facility input transfers T_{in}, output transfers T_{out}, and inventory I to compute a material balance defined for balance period j as $\mathrm{MB}_j = (\mathrm{I}_{j-1} + \mathrm{T}_{\mathrm{in,j}} - \mathrm{T}_{\mathrm{out,j}}) - \mathrm{I}_j = b$ook inventory-physical inventory, where $(\mathrm{I}_{j-1} + \mathrm{T}_{\mathrm{in,j}} - \mathrm{T}_{\mathrm{out,j}})$ is the book inventory. Typically, many measurements are combined to estimate the terms $\mathrm{T_{in}}$, $\mathrm{I_{begin}}$, $\mathrm{T_{out}}$, and $\mathrm{I_{end}}$ in the MB; therefore, the central limit effect and years of experience suggests that MBs in most facilities will be approximately normally distributed with mean equal to the true SNM loss μ and standard deviation σ_{MB}, which is expressed as $\mathrm{X} \sim \mathrm{N}(\mu, \sigma_{\mathrm{MB}})$, where X denotes the MB [5,14]. Therefore, a sequence of n MBs are assumed to have approximately a multivariate normal distribution, $X_1, X_1, \ldots,$ $X_1 \sim \mathrm{MVN}(\mu, \Sigma)$ where the n-by-n covariance matrix $\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 \ldots \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \ldots . \sigma_{2n}^2 \\ \cdots & \\ \sigma_{n1}^2 & \sigma_{n2}^2 \ldots . \sigma_{nn}^2 \end{pmatrix}$.

The magnitude of σ_{MB} determines what SNM loss μ leads to high detection probability (DP). For example, suppose the facility tests for SNM loss only, not for SNM gain, and assume that $\mathrm{X} \sim \mathrm{N}(0, \sigma_{\mathrm{MB}})$ is an adequate model. Then, if a false alarm probability of σ_{MB}= 0.05 is desired, the alarm threshold is 1.65 σ_{MB}. Or, if the facility tests for loss or gain, then the alarm threshold is 1.96 σ_{MB}In the case of testing for loss only, it then follows that the loss detection probability 1- β for μ = 3.3 $\sigma_{\mathrm{MB}}$and 1 – β > 0.95 if μ > 3.3σ_{MB}, where β is the nondetection (false negative) probability. The factor 3.3 arises from symmetry of the Gaussian, requiring FAP = α= 0.05, and the fact that 1.65=3.3/2 is the 0.95 quantile of the N(0,1) distribution. If the facility tests for loss or gain, then 1 – β > 0.95 if μ > (1.65+1.96) σ_{MB}=3.61σ_{MB}B . The DP of other safeguards measures such as enhanced containment and surveillance with smart cameras and/or remote radiation detection is difficult to quantify and is outside the scope.
One common goal is for the loss detection probability DP=$1 - \beta$ to be at least 0.95 if $\mu \geq 1$ SQ (significant quantity, which is 8 kg for Pu), which is accomplished if $\sigma_{\mathrm{MB}} \leq$ SQ/3.3. If σ_{MB} > SQ/3.3, this can be mitigated either by reducing the typical

magnitude of measurement errors to achieve $\sigma_{MB} \leq$ SQ/3.3 (if feasible), and/or by closing the balances more frequently so there is less nuclear material transferred per balance period, which reduces σ_{MB}. It is important to recognize that large throughput facilities cannot typically achieve DP ≥ 0.95 for a loss of $\mu \geq 1$ SQ over a long time period such as one year. And, NRTA is not a panacea, because, as shown in [ref], if a facility slowly diverts NM over, for example, one year, then a single yearly statistical test has larger DP than frequent statistical testing during the year. Of course if the facility diverts NM abruptly, such as over one day, then NRTA will have much larger DP than a single annual statistical test. It is therefore generally accepted that NRTA is a valuable safeguards measure, despite leading to slightly smaller DP than in using annual MBs for protracted loss detection. Most safeguards studies assume that a yearly decision is made, corresponding to the time of the annual scheduled physical inventory. But, if the facility diverts, for example, SQ/2 in year 1 and SQ/2 in year 2, then the DP is lowered; however, the diversion time is longer than one year. See Figure 6; however, the required diversion time would be longer than one calendar year, in this figure, lasting from period 7 to 18.

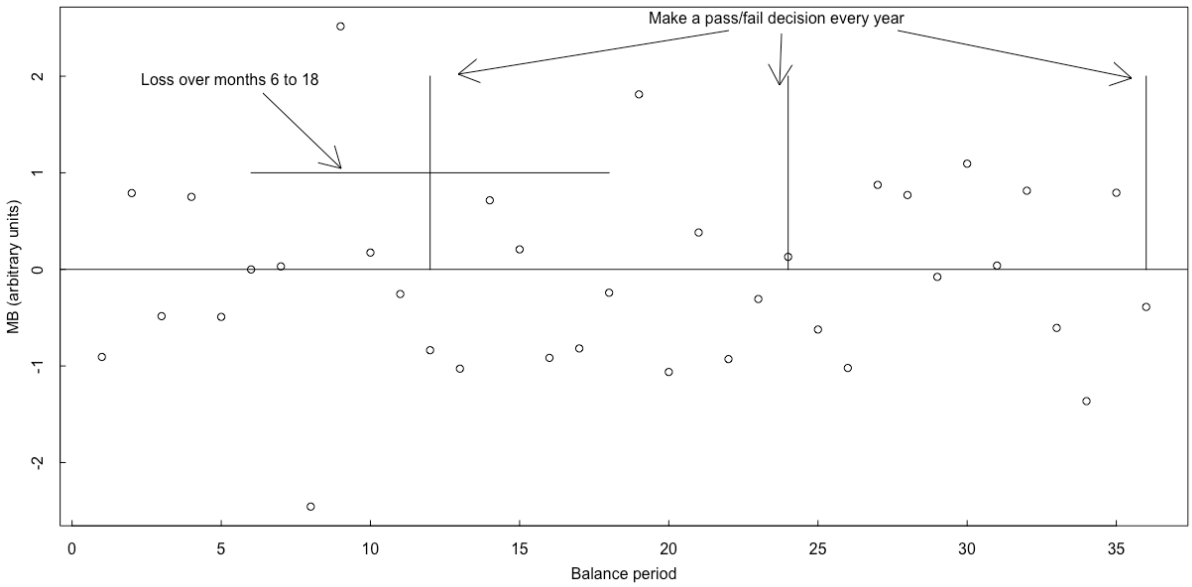

Figure 6. MB sequences over 36 months using fixed-period (annual) decision periods.

4.1 Propagation of Variance to Estimate Σ

The following is a simplified example [14, 43] of estimating Σ using a model of a generic electrochemical facility with one input stream, one output stream, and two inventory items is as follows. First, each individual measurement method is modelled with a measurement error model. A typical model for multiplicative errors for the operator (O) is $O_i = \mu_i(1 + S_{Oi} + R_{Oi})$ with $S_{Oi} \sim N(0, \delta_{SO}^2)$ and $R_{Oi} \sim N(0, \delta_{RO}^2)$, where O_i is the operator's measured value of item i, μ_i is the true but unknown value of item i, R_{Oi} is a random error of item i, $S_{Oi} + R_{Oi}$is a short-term systematic error for item i. Then, the diagonal terms of Σ are calculated as

$$\sigma_i^2 = in_i^2\left(\delta_{in,R}^2 + \delta_{in,S}^2\right) + out_i^2\left(\delta_{out,R}^2 + \delta_{out,S}^2\right) +$$

$$\sum_{k=1}^{2}\left\{inv_i^2\left(\delta_{inv,R}^2 + \delta_{inv,S}^2\right) + inv_{i-1}^2\left(\delta_{inv,R}^2 + \delta_{inv,S}^2\right)\right\} - 2\sum_{k=1}^{2}\left(inv_i inv_{i-1}\delta_{inv,S}^2\right)$$

The off-diagonal terms Σ are calculated as:

$$\sigma_{ij}^2 = in_i in_j\delta_{in,S}^2 + out_i out_j\delta_{out,S}^2 + \sum_{k=1}^{2}\left\{\left(inv_i inv_j + inv_{i-1}inv_{j-1}\right)\delta_{inv,S}^2\right\} -$$

$$\sum_{k=1}^{2} inv_i inv_{j-1}\left(\delta_{inv,S}^2 + \delta_{inv,R}^2\left[if\ j-i=1\right]\right) - \sum_{k=1}^{2} inv_{i-1}inv_j\left(\delta_{inv,S}^2 + \delta_{inv,R}^2\left[if\ i-j=1\right]\right)$$

In the last two terms, the random error of the inventory term is only applied if the condition is true.

4.2 Sequential Testing

The assumption $X_1\ X_2, \ldots, X_N \sim \text{MVN}(\mu, \Sigma)$ implies that $Y = \Sigma^{-1/2}X \sim MVN\left(\Sigma^{-\frac{1}{2}}\mu, I\right)$, where *I* is the identity matrix. The transform $Y = \Sigma^{-1/2}X$ is known in safeguards as the standardized-independently-transformed MUF (SITMUF, where MUF is another name for the MB), which is most conveniently computed using the Cholesky decomposition [14,20,43]. There are two main advantages of applying statistical tests to *Y* rather than to *X*. First, alarm thresholds depend only on the sequence length *n* for *Y*, and not on the form of the covariance matrix Σ. Because it is best to calculate thresholds using simulation, this is a logistic advantage. Second, the variance of the *Y* sequence decreases over time, so particularly if a diversion occurs late in the analysis period, the DP is larger for the *Y* sequence than for the *X* sequence. Note that one cannot claim higher DP for the *Y* sequence than for the *X* sequence in general, because the true loss scenario is never known, and the DP can be larger for *X* than for *Y* for some loss scenarios, which is demonstrated in Section 4.

The value of Y_i can be calculated using $Y = \Sigma^{-1/2}X$, but more intuitively as the residual from the *X* sequence, $Y_i = X_i - E(X_i|X_1\ X_2, \ldots, X_{i-1})$ where *E* denotes the expectation and the standard deviation σ_i is given by $\sigma_i = \sqrt{\sigma_{ii}^2 - f\Sigma^{-1}f^T}$ where $f = \Sigma_{i,1:(i-1)}$, the 1 to (*i*-1) entries in the *i*th row of Σ.

Several reasonable statistical tests have been evaluated in [14,19,23,43], and are included in the simulation study in Section 4, including:

4.2.1 MUF test. This compares each *x* = MUF value to a threshold, which is the same as a Shewhart test. The test alarms on period I if $\frac{MUF_i}{\sigma_{MUF,i}} \geq T$ for some threshold T.

4.2.2 SITMUF test. This compares each SITMUF value to a threshold, which is the same as a Shewhart=1 test in QC. The test alarms on period *i* if $\frac{ITMUF_i}{\sigma_{SITMUF,i}} \geq T$ for for some threshold *T*, where $\sigma_{SITMUF_i} = 1$.

4.2.3 Page applied to MUF. Page's test [14,15,19,43,44] to test for loss is a sequence of sequential probability ratio tests, defined as $P_i = \max(0, P_{i-1} + \frac{x_i}{\sigma_i} - k)$. Alarm on period *t* if $P_t > h$. The parameter k is a control parameter that is optimal for detecting a shift from zero loss to loss L if $k = \frac{L}{2}$. The alarm threshold h is chosen so that the FAP per analysis period (usually one year) is 0.05 or whatever FAP is specified.

4.2.4 Page applied to SITMUF. Page's test to test for loss is a sequence of sequential probability ratio tests, defined as$P_i = \max(0, P_{i-1} + xy_i - k)$.. The alarm threshold h is chosen so that the FAP per analysis period (usually one year) is 0.05 or whatever FAP is specified.
Page's test with a large value of k and small value of h has good DP for abrupt loss, and with a small value of k and large value of h has good DP for protracted loss. Therefore, a reasonable option is to use a combination of two Page's tests, one with large k and one with small k.
4.2.5 Apply combined Page's tests to MUF. Choose a relatively large value of k and small value of h to tune Page's test to have large DP for abrupt loss, or a small value of k and large value of h to have large DP for abrupt loss. Use the combination as a test.
4.2.6 Same as 4.2.5, but apply to SITMUF.
4.2.7 Sequential CUMUF. At period i, $\text{CUMUF}_i = \sum_{j=1}^{1} x_j$ is the sum of all MUF values from period 1 to i.
4.2.8 GEMUF. It has been shown that if the loss vector μ is known, then the Neyman-Pearson test statistic is $\mu^T\Sigma^{-1}x$, which is known as a matched filter in some literature. The GUMUF statistic substitutes x^T for μ^T, so $GEMUF = x^T\Sigma^{-1}x$. In simulation studies, μ is known, so the NP test statistic is useful for calculating the largest possible DP. Geschatzter is "estimated MUF" in german, and GEMUF is the same as the Mahalanobis distance from the 0 vector, and Hotelling's multivariate T statistic, which are used in multivariate process control
4.2.9 A nonsequential version of the Neyman-Pearson test, $\mu^T\Sigma^{-1}x$, is useful to calculate the largest possible DP for a given Σ and μ.
4.2.10 A nonsequential CUMUF (the annual CUMUF), which is often included in a suite of tests.

The SITMUF transform is recommended for two reasons. First, simulation is typically used to select alarm thresholds, and it is convenient to always work on the same scale when selecting alarm thresholds, so the fact that $Y = \Sigma^{-1}X \sim \text{MVN}(\Sigma^{-\frac{1}{2}}\mu, I)$ is convenient. Note that alarm thresholds could be selected on the basis of exact or approximate analytical results for some, but not all, of the tests. For example, there are approximate expressions for h and k (Brook and Evans). Second, the standard deviation σ_i is given by $\sigma_i = \sqrt{\sigma_{ii}^2 - f\Sigma^{-1}f^T}$where $f = \Sigma_{i,1:(i-1)}$, is the 1 to (i-1) entries in the ith row of Σ, so the standard deviation of the MUF residuals decreases in the later periods. Therefore, the independence transform is analogous to a bias adjustment, leading to smaller prediction variance in later periods, which tends to increase the DP for SITMUF compared to MUF (there are exceptions where the DP for MUF is larger than the DP for SITMUF; see Section 4 results).
Remark 4.1. Thresholds can be chosen in many ways [13,14,18,20,43-45], and can be assumed to be constant for each period or not.
Remark 4.2. Performance criteria. The main performance criterion for comparing tests is the DP. But, the average time to detection and robustness to misspecifying the covariance matrix Σ are also important.
Remark 4.3. There are other tests in the literature [16,17,23]. A few other tests have been proposed for NRTA, including the power one test and the scan statistic. The power one test is not relevant in the context of fixed-period decision making.
Remark 4.4. Several simulation studies have been published [14,15,19,23,43].

4.3 Hybrid testing of NM and process monitoring data

Options to quantify the benefit of PM data by using P(alarm|diversion scenario) as the figure of merit are described in [15], while using both PM and NMA residuals in the alarm rule. A key assumption is that the safeguards approach includes model-based predictions that can be compared to corresponding measurements, resulting in time series of residuals. The requirement for high-quality predictions leads to technical challenges in safeguarding either aqueous or electrochemical reprocessing facilities. For example, there is ongoing work aimed at high-quality modeling of the electrorefiner in an electrochemical facility [14, 43]. Strictly speaking, our approach leads to high SNM loss detection probability only for the specified diversion routes; however, there is also high loss detection probability for any type of abrupt loss.. Also, in the context of international safeguards, there is not yet an approach to authenticate operator PM data; authentication will depend on facility type and is under investigation.

In general, we propose to estimate the DP of the safeguards system by estimating the system DP from PM combined with NMA using the following two steps:

a) Describe diversion scenarios to inform how PM data should be evaluated to provide a means of event detection, and
b) Evaluate P(alarm|diversion scenario), the conditional probability of an alarm for a given scenario. The alarm rule operates on p residuals r_1, r_2, …., r_p which include MB values from NMA, plus residuals from monitoring "wait" and "transfer" modes in tank SM data. The probability P(alarm|diversion scenario) is a function of the true states of nature, the measurement system, and the alarm rule.

To illustrate, Figure 7 plots example PM data from which residuals (measurement minus prediction) can be calculated [15]. Figure 8 plots PM residuals and the MBs from NMA. Figure 8 plots example DPs from the maximum of all the Page's cusums and the average of all the Page's cusums for (a) a localized in time and space loss, and (b) a non-localized in time and space loss.

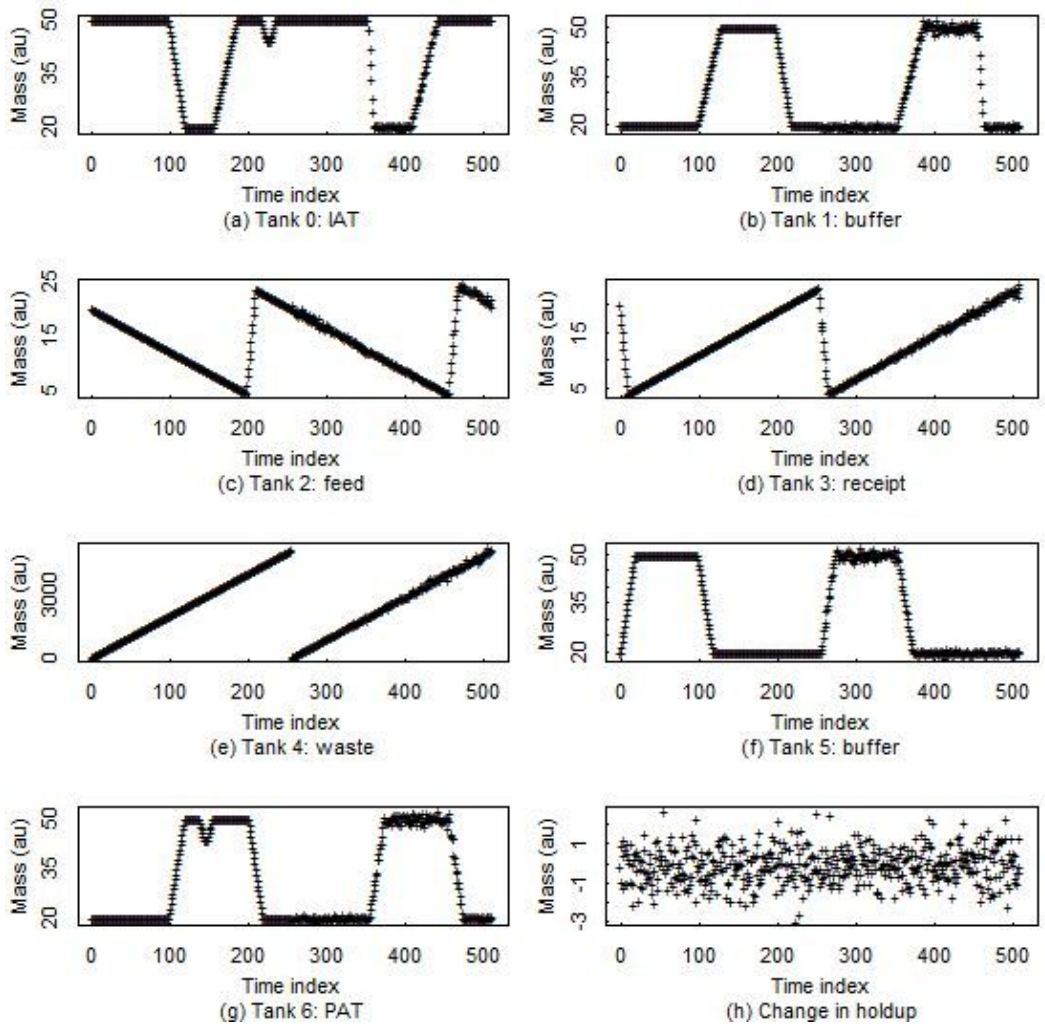

Figure 7. Process monitoring data from an input accountability tank (IAT), buffer1 tank, feed tank, receipt tank, waste, buffer2 tank, product accountability tank, and from comparing predicted to measured holdup (material that is in the process in difficult-to-measure form).

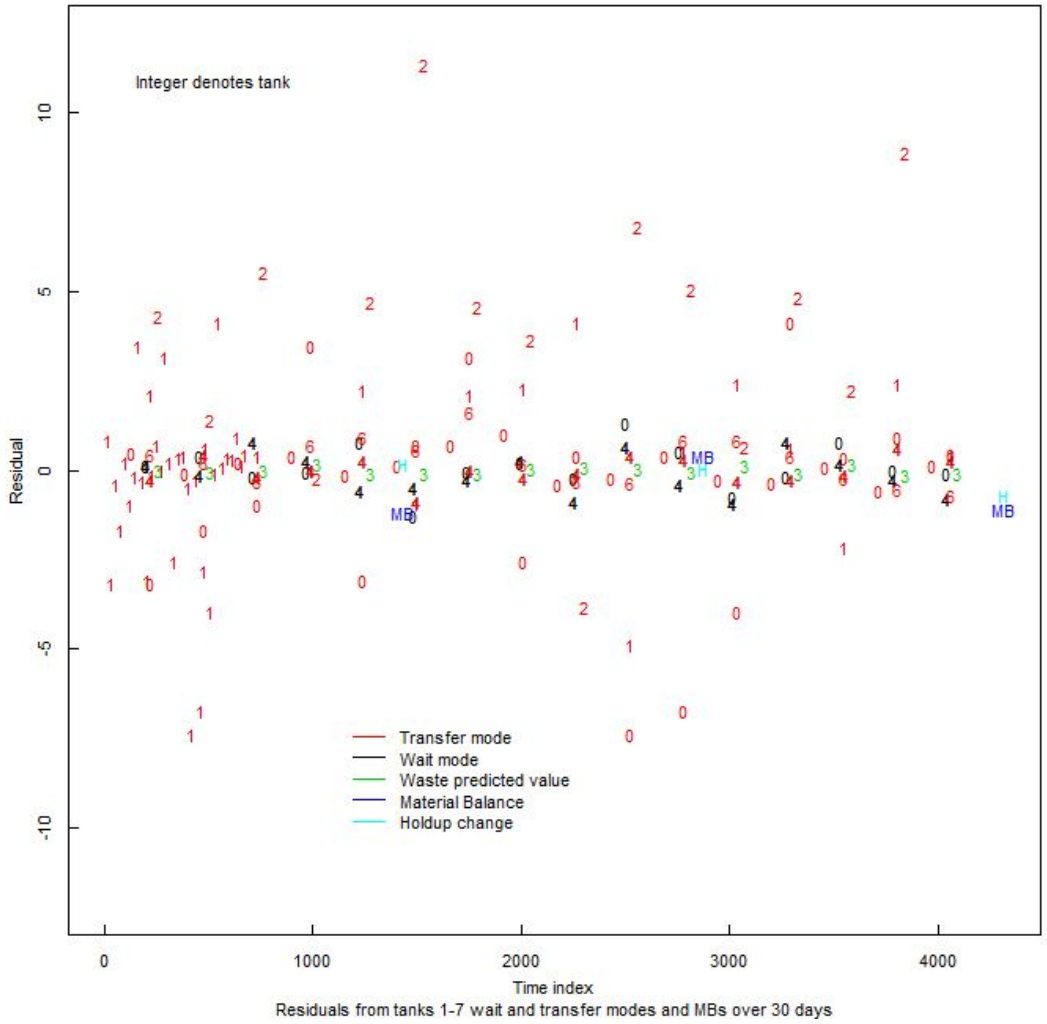

Figure 8. Example simulated MB and PM data to which sequential testing can be applied.

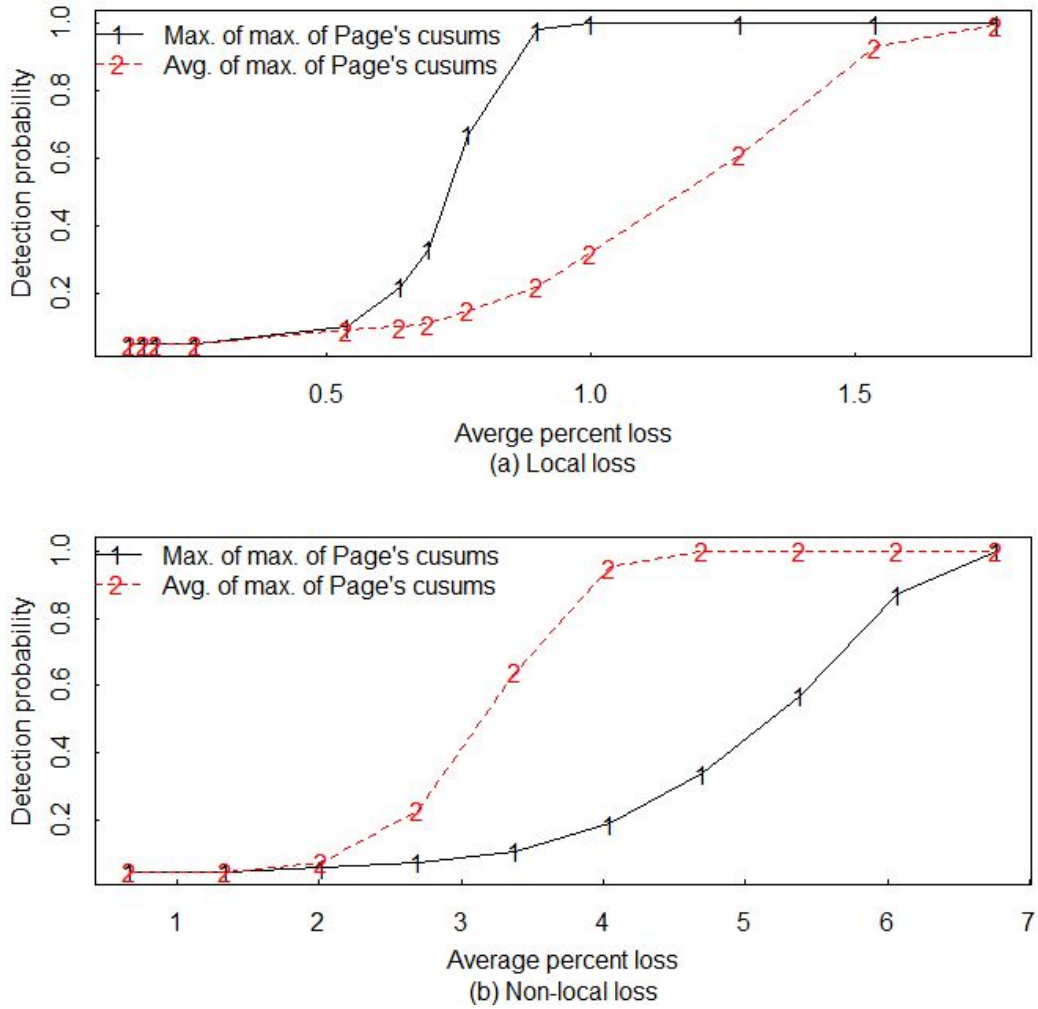

Figure 9. Example DPs from the maximum of all the Page's cusums and the average of all the Page's cusums for (a) a localized in time and space loss, and (b) a non-localized in time and space loss.

4. Summary

This paper reviewed statistical methods for UQ of measurements, for constructing tolerance intervals for setting pass/fail criteria for monitored data streams, and for estimating DPs for specified NM misuse scenarios at declared facilities. UQ for measurements was done both empirically using data collected for metrology studies and from applying error variance propagation to all steps in the assay (physics-based). Approximate Bayesian computation (ABC) was used for both the empirical and physics-based UQ. The estimated measurement error RSDs were then used to estimate the SD of the NM mass balances that are analyzed sequentially over time.

References

1. Bonner, E., Burr, T., Guzzardo, T., Norman, C., Zhao, K., Beddingfield, D., Geist, W. Laughter, M., Lee, T., Ensuring the effectiveness of safeguards through comprehensive uncertainty quantification, Journal of Nuclear Material Management 44, 53–61, 2016.
2. 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 Material Management 45, 4–14. 2017.
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 of Nuclear Science and Technology *2,* 36*,* 2016.
4. Burr, T., Bonner, E., Krzysztoszek, K., Norman, C., Setting alarm thresholds in measurements with systematic and random errors, Journal of Open Statistics *2*, 259–271, 2019.
5. Burr, T., Hamada, M.S., Smoothing and time series modeling of nuclear material accounting data for protracted diversion detection, Nuclear Science and Engineering 177, 307–320, 2014.
6. Martin, K., Böckenhoff, A., Analysis of short-term systematic measurement error variance for the difference of paired data without repetition of measurement. Advances in Statistical Analysis 91, 291–310, 2007.
7. 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..
8. 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.
9. Zhao, K., Penkin, P., Norman, C., Balsely, S., Mayer, K., Peerani, P., Pietri, P., Tapodi, S., Tsutaki, Y., Boella, M., et al. STR-368 International target values 2010 for measurement uncertainties in safeguarding nuclear materials, IAEA, Vienna, 2010. Available online: www.inmm.org (accessed on 12/17/201712).
10. Agboraw, E. Bonner, E., Burr, T., Croft, S., Kirkpatrick, J., Krieger, T., Norman, C., Santi, P., Revisiting Currie's minimum detectable activity for nondestructive assay by gamma detection using tolerance intervals. ESARDA Bulletin 54, 14–22, 2017.
11. Carlin, B., John, B., Stern, H., Rubin, D., Bayesian Data Analysis, 1st ed.; Chapman and Hall: Boca Raton, Fla., 1995.
12. Avenhaus, R., Canty, M., Compliance Quantified. An Introduction to data verification, Cambridge Press, Cambridge, 1977.
13. Avenhaus, R., Jaech, J., On subdividing material balances in time and/or space, Journal of Nuclear Materials Management 10(3), 24–33, 1981.
14. 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, 2013. doi:10.1155/2013/961360, 2013.
15. Burr, T., Hamada, M.S., Skurikhin, M., Weaver, B., Pattern recognition options to combine process monitoring and material accounting data in nuclear safeguards, Statistics Research Letters 1(1), 6-31, 2012.
16. Downing, D. J., Pike, D. H., Morrison, G. W. ,Analysis of MUF data using ARIMA models, Journal of Nuclear Materials Management 7, 80–86, 1978.

17. Downing, D. J., Pike, D. H., Morrison, G. W., Application of the Kalman Filter to Inventory Control, Technometrics 22, 17–22, 1980.
18. Goldman, A. S., Picard, R. R., Shipley, J. P., Statistical methods for nuclear material safeguards, Technometrics 24, 267–275, 1982.
19. Jones, B., Calculation of diversion detection using the SITMUF sequence and Page's test: application to evaluation of facility designs, in Proceedings of the 7th ESARDA Symposium on Safeguards and Nuclear Material Management, Liege, Belgium, 1985.
20. Picard, R., Sequential analysis of materials balances, Journal of Nuclear Materials Management, 15(2), 38–42, 1987.
21. Sanborn, J. A method for assessing safeguards effectiveness and its application to state-level material accountancy verification, Journal of Nuclear Materials Management 43(4), 19-33, 2015.
22. Sher, R., Untermyer, S., The Detection of Fissionable Material by Nondestructive Means. American Nuclear Society, LaGrange Park, IL, 1980.
23. 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.
24. Willrich, M., International Safeguards and Nuclear Industry. John Hopkins University Press, Baltimore, 1973.
25. Burr, T., Martin, K., Norman., C., Zhao, K., Analysis of variance for item differences in verification data with unknown groups. Science and Technology of Nuclear Installations 1769149, doi:10.1155/2019/1769149, 2019.
26. Thompson, M. and Ellison, S., Dark uncertainty, Accreditation and Quality Assurance 16, 483– 487, 2011
27. Guide to the Expression of Uncertainty in Measurement, JCGM 100, www.bipm.org, 2008.
28. ASTM C1514: Standard Test Method for Measurement of ^{235}U Fraction using the Enrichment Meter Principle, 2008.
29. Bich, W., Revision of the 'guide to the expression of uncertainty in mea- surement', why and how. Metrologia 51, S155–S158, 2014.
30. Miller, R., Beyond ANOVA, Chapman and Hall, 1998.
31. Carlin, B., John, B., Stern, H., Rubin, D., Bayesian Data Analysis, 1st ed.; Chapman and Hall: Boca Raton, Fla., 1995.
32. Fearnhead, P., Prangle, D., Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation, Journal of the Royal Statistical Society B 74, 419–474, 2012.
33. Joyce, P., Marjoram, P. Approximately sufficient statistics and Bayesian computation. Statistical Applications in Genetics and Molecular Biol*ogy 7*, 26-44, 2008.
34. Burr, T., Skurikhin, A., Selecting summary statistics in approximate Bayesian computation for calibrating stochastic models. Biomed. Res. Int. *2013*, 210646, doi:10.1155/2013/210646, 2013.
35. Blum, M., Nunes, M., Prangle, D., Sisson, S. A comparative review of dimension reduction methods in approximate Bayesian computation. Statistical Science *28*, 189–208, 2013.
36. Nunes, M., Prangle, D., abctools: An R package for tuning approximate Bayesian computation analyses. The R Journal *7*, 189–205, 2015.

37. R Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2012; ISBN 3-900051-07-0. Available online: http://www.R-project.org (02/17/2015).
38. Burr, T., Krieger, T., Norman, C., Approximate Bayesian computation applied to metrology for nuclear safeguards ESARDA bulletin 57, 50-59, 2018.
39. Burr, T., Knepper, P., A study of the effect of measurement error in predictor variables in nondestructive assay, Applied Radiation and Isotopes 53, 547–555, 2000.
40. Kang, P, Koo, C., Roh, H., Reversed inverse regression for the univariate linear calibration and its statistical properties derived using a new methodology, International Journal of Metrology and Quality Engineering 8 (28) 1-10, 2017.
41. Kraemer, K. Confidence interval for variance components and functions of variance components in the random effects model under non-normality. Ph.D. Thesis, Iowa State University, Ames, IA, USA, 2012.
42. Burr, D. bspmma: An R package for Bayesian semiparametric models for meta-analysis. Journal of Statistical Software 50, 1–23, 2012.
43. Won, B., Shin, H., Park, S., Ahn, S., Development of PYMUS+ code for quantitative evaluation of nuclear material accounting system, Science and Technology of Nuclear Installations Vol 2019, Article ID 8479181, 11 pages, https://doi.org/10.1155/2019/ 8479181, 2019.
44. Page, E., Continuous inspection schemes, Biometrika 41, 100–115, 1954.
45. Brook, D., Evans, D., An approach to the probability distribution of cusum run length, Biometrika 59 (3), 539-549, 1972.