

# Bayesian Updating of Material Balances Covariance Matrices Using Training Data

T. Burr<sup>1</sup>, M.S. Hamada<sup>2</sup>

<sup>1,2</sup> *Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos NM 87545, USA*

*tburr@lanl.gov*

*hamada@lanl.gov*

## ABSTRACT

The main quantitative measure of nuclear safeguards effectiveness is nuclear material accounting (NMA), which consists of sequences of measured material balances that should be close to zero if there is no loss of special nuclear material such as Pu. NMA is essentially “accounting with measurement errors,” which arise from good, but imperfect, measurements. The covariance matrix  $\Sigma_{MB}$  of a sequence of material balances is the key quantity that determines the probability to detect loss. There is a recent push to include process monitoring (PM) data along with material balances from NMA in new schemes to monitor for material loss. PM data includes near-real-time measurements by the operator to monitor and control process operations. One concern regarding PM data is the need to estimate normal behavior of PM residuals, which requires a training period prior to ongoing testing for material loss. Assuming that a training period is used for PM data prior to its use in statistical testing for loss, that same training period could also be used for improving the estimate of  $\Sigma_{MB}$  that is used in NMA. We consider updating  $\Sigma_{MB}$  using training data with a Bayesian approach. A simulation study assesses the improvement gained with larger amounts of training data.

## 1. INTRODUCTION

Nuclear material accounting (NMA) at safeguarded facilities consists of periodic measurement of special nuclear material (SNM) flows and inventories, both of which are measured with nonnegligible errors. Typically, the SNM mass is measured using a combination of techniques. For example, an aqueous reprocessing facility chops and dissolves spent fuel, then purifies the recovered Pu in a series of separation steps. Such a facility recovers, for example, Pu from spent fuel, and the main measurements are bulk solution volume and Pu concentration. The Pu mass is then estimated by multiply-

ing measured volume by measured Pu concentration. NMA compares the measured Pu in inputs, outputs, and inventory to track Pu mass during facility operations.

The main quantitative assessment is the material balance (MB), defined as  $MB = \text{beginning inventory} + \text{transfers in} - \text{transfers out} - \text{ending inventory}$ . Metrology for each measurement method, such as bulk weight and SNM concentration in a tank, provide estimates of individual measurement errors in each inventory and transfer term. Propagation of variance is used to estimate the covariance  $\Sigma_{MB}$  of a sequence of MBs by combining the effects of individual “random” and “systematic” errors (Goldman et al., 1992; Aigner et al., 2010; Beedgen, 1987). Random errors impact the measurement of only one item. Systematic errors impact the measurements of multiple items; therefore, while systematic errors can cancel or partly cancel out in some cases (for example, when comparing two volume measurements in a tank), in other cases they have a large impact on  $\Sigma_{MB}$ . Multiple items can share a systematic error due to calibration effects, leading to a systematic positive or negative error until the assay method is recalibrated.

Time series modeling and various forms of data mining, such as nonparametric smoothing, have been proposed in the NMA context in response to the fact that the loss detection probability of small sustained (“protracted”) loss of SNM over multiple time periods is lower than desired (Prasad et al., 1995a, 1995b; Hamburg et al., 1996; Grznar et al., 1997; Downing et al., 1978). Such data mining requires up-front training data prior to testing for SNM loss. Assuming that MB sequences have approximately a multivariate normal distribution, with known covariance matrix  $\Sigma_{MB}$ , Burr and Hamada (2014a) show that traditional parametric modeling of the MB sequences leads to higher detection probability than nonparametric modeling. The results in Burr and Hamada (2014a) confirm that parametric modeling performs better than nonparametric modeling if the parametric modeling assumptions are true or close to true. One contribution of Burr and Hamada (2014a) is to clarify that protracted loss over multiple time pe-

Tom Burr et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

riods is simply difficult to detect; and, under the assumption that the MB sequence is multivariate normal, an option that exploits the assumed multivariate normality has the highest possible detection probability for any type of SNM loss. Note that the multivariate normal assumption,  $MB \sim N(0, \Sigma_{MB})$  if there is zero loss, is a reasonable assumption because of the central limit theorem applied to the many measurements that are combined in typical MB calculations (Downing et al., 1978; Burr and Hamada, 2013a; Jones, 1989). However, there is always some estimation error in  $\Sigma_{MB}$ , that Burr and Hamada (2013a) showed can have non-negligible impact on the false alarm probability and on the detection probability at a given facility. The focus of this paper is to evaluate to what extent one can reduce the estimation error in  $\Sigma_{MB}$  by using training data, as illustrated below.

Partly in response to low detection probabilities for protracted loss, there is recent interest in extending quantitative assessments to include frequently recorded residuals from process monitoring (PM) in addition to less frequent MB values (Burr et al. 2012, 2013a, 2013b). PM residuals include in-vessel readings and comparisons of data to various engineering estimates and models. A disadvantage of such PM is the need for training data to establish alarm thresholds prior to testing for SNM loss (Burr et al., 2013a, 2014b). Also, Avenhaus and Jaech (1981) show that analyzing more frequent MBs actually decreases loss detection probability for protracted loss, due to requiring high alarm thresholds to account for more statistical tests per unit time. Although more frequent MBs will slightly decrease the loss detection probability for worst-case protracted loss, specialized PM and high-frequency MBs can have high loss detection probability for specified loss scenarios, while only slightly decreasing the loss detection probability for the worst-case protracted loss (Burr et al., 2012).

Quantitative evaluations of the loss detection probability for sequential tests of MB sequences with various loss scenarios have been presented (Beedgen, 1987; Avenhaus and Jaech, 1981; Jones, 1989), but estimation error in  $\Sigma_{MB}$  was ignored. The most commonly-used sequential test in safeguards is Page's cusum test applied to a transform of the MB sequence (known as the SITMUF sequence, the standardized independently transformed MUF sequence, where MUF = MB is the international term "material unaccounted for"). Options are being developed to use both PM residuals and NMA sequences in multivariate statistical process monitoring (Burr et al., 2012), some of which use multivariate versions of Page's cusum test.

As mentioned above, the requirement for training data in the PM context was proposed in previous literature, but in the context of training methods to predict MB values in testing. The main topic of this paper is the potential to use training data in order to improve the estimate of  $\Sigma_{MB}$ , which has not yet been considered. It is assumed that both PM and NMA

can use training data prior to testing for SNM loss. In the NMA context, we illustrate how to use such training data to improve the estimate of  $\Sigma_{MB}$ .

The current emphasis on quantitative roles for PM leads to a new challenge for domestic safeguards. The challenge is to simultaneously monitor for changing sensor health and for SNM loss using relatively frequent monitoring. Traditionally, this journal does not deal with the topic of monitoring for material loss, but deals extensively with monitoring sensor health. In measurement control for safeguards, one typically uses frequent consistency checks to monitor sensor health. For example, known volume additions can check instrument responses for stability. Also, tank-to-tank transfers can be monitored for consistency; however, if one aims to simultaneously monitor for either SNM loss or instrument anomalies or drift, then some type of redundancy checks are necessary. Such instrument health checks lead to confidence that the infrequently-reported MB values are not subject to instrument-related artifacts. International safeguards has the same challenge as domestic safeguards, but the possibility of data falsification by the operator to conceal SNM diversion (a type of loss) is an added complication. Howell et al. (2013) present a semi-quantitative approach for domestic safeguards to simultaneously monitor for changing sensor health and SNM loss. The approach assumes that some sensor redundancy is available, and that neighboring sensors will not simultaneously malfunction. To our knowledge, there has not been an analogous quantitative attempt for international safeguards.

This article focuses on the challenge of monitoring for SNM loss, assuming that sensor health is monitored using measurement control data. The article considers updating an estimated  $\Sigma_{MB}$  using training data with a Bayesian approach. We assume that an initial estimate of  $\Sigma_{MB}$  is available, as usual, from metrology data that can include calibration data, measurement control data, and perhaps, measurement comparison data, for each type of measurement that is used in the MB calculation. While PM requires training data prior to monitoring for SNM loss, in principle, NMA does not, provided the metrology data leads to an adequate estimate of  $\Sigma_{MB}$ . In practice, such estimates are often based on very limited data and can be revised as more meta data become available. So, in practice, NMA is vulnerable to having a poorly estimated  $\Sigma_{MB}$  in the early history of an operating facility, and in effect, also relies on a training period prior to quantitative monitoring for SNM loss.

An outline is as follows. Section 2 provides additional background. Section 3 provides details of the form of  $\Sigma_{MB}$  for five cases depending on whether there is systematic measurement error, whether there is an end-of-year cleanout, and whether there is an end-of-year measurement calibration. Section 4 briefly reviews Bayesian inference. In Section 5, we outline

a simulation study to assess the improvement gained by using training data to update  $\Sigma_{MB}$ . Section 6 reports results of the simulation study. Section 7 describes a second option for analysis of the training data. Section 8 describes sequential tests in use for NMA and what the simulation study results imply about their performance in the presence of estimation error in  $\Sigma_{MB}$ . We conclude with a discussion in Section 9.

## 2. BACKGROUND

Large facilities are divided into material balance areas, such as the first separation/purification area of an aqueous reprocessing facility. The area includes feed, receipt, buffer, and storage tanks that all handle Pu solution, as well as a waste stream and a chemical separations area that holds Pu, but is not equipped for Pu measurement. Approximately every 10 days, a MB is computed, usually without cleaning out the holdup in the separations area, pumps, pipes, or in the uninstrumented tanks. The MB definition implies that if there is SNM loss, then the expected value of the MB is positive; if there is SNM gain, then the expected value of the MB is negative. Statistical tests analyze sequences of MBs for SNM loss as evidenced by one or more large positive MBs that is (are) too large to have been likely to arise from measurement errors.

In a typical aqueous reprocessing plant, the Pu mass is estimated as  $M = VC$ , where  $V$  is the bulk volume measurement, and  $C$  is the Pu concentration measurement. Following international target values for uncertainty (Aigner et al., 2010), we assume  $V = V_{\text{true}}(1 + S_V + R_V)$  and  $C = C_{\text{true}}(1 + S_C + R_C)$ , which collectively imply that the measurement equation for  $M$  is  $M \approx M_{\text{true}}(1 + S_V + R_V + S_C + R_C)$ , ignoring products of errors, because they are negligible. The net variance for the Pu mass in one tank is therefore approximately  $M_{\text{true}}^2(\tilde{\sigma}_{S,M}^2 + \tilde{\sigma}_{R,M}^2)$ , where  $\tilde{\sigma}_{S,M}^2 = \tilde{\sigma}_{S,V}^2 + \tilde{\sigma}_{S,C}^2$  and  $\tilde{\sigma}_{R,M}^2 = \tilde{\sigma}_{R,V}^2 + \tilde{\sigma}_{R,C}^2$ , and here we use the tilde notation for relative error variances. Note that  $MB_i = \sum_{i=1}^{nin_i} (V_i C_i) - \sum_{i=1}^{nout_i} (V_i C_i) + \sum_{i=1}^{ninv_{i-1}} (V_i C_i) - \sum_{i=1}^{ninv_i} (V_i C_i)$ , so additional variance propagation is used to estimate the covariance matrix  $\Sigma_{MB}$  of a sequence of MBs (see Section 3).

A key fact from the variance propagation is that  $\Sigma_{MB}$  depends on the true inputs, outputs, and inventories, and on only a few aggregate variance components, ( $\tilde{\sigma}_{S,M}^2$  and  $\tilde{\sigma}_{R,M}^2$ ). The true inputs, outputs, and inventories are estimated well enough for variance propagation purposes by the corresponding measurements (because the overall standard deviation of an MB is relatively small, typically 1 % or less of the SNM throughput). Although usually measured with large relative uncertainty, material holdup in pumps, pipes, and un-instrumented areas, can be included in the inventory terms. The aggregate variance components should be somewhat well estimated from auxiliary measurement control data and measurement

studies. The main point of this paper is to illustrate that MB data provides an additional data source to improve the initial estimates of  $\tilde{\sigma}_{S,M}^2 = \tilde{\sigma}_{S,V}^2 + \tilde{\sigma}_{S,C}^2$  and  $\tilde{\sigma}_{R,M}^2 = \tilde{\sigma}_{R,V}^2 + \tilde{\sigma}_{R,C}^2$ , which leads to improved estimation of  $\Sigma_{MB}$ . As we demonstrate in Section 8, improved estimation of  $\Sigma_{MB}$  means that sequential tests (which use the estimate of  $\Sigma_{MB}$ ) have actual false alarm rates that are closer to the nominal false alarm rates (and the actual detection probabilities for a specified loss of SNM are closer to the nominal detection probabilities).

## 3. COVARIANCE MATRICES FOR MB DATA

Burr and Hamada (2013a) considered five covariance matrices for MB data over two years. Using the same type of measurement error equation as in Section 2, and the same simple variance propagation, one can calculate the covariance matrix  $\Sigma_{MB}$  of a sequence of MBs. The five covariances correspond to five distinct scenarios. Scenario one has no cleanout, and no systematic measurement error (as in Avenhaus and Jaech, 1981). Scenarios two to five each have systematic error. Scenario two has no cleanout, with measurement calibration. Scenario three has no cleanout, without measurement calibration. Scenario four has cleanout, with measurement calibration. Scenario five has with cleanout, without measurement calibration. These covariance matrices are denoted  $\Sigma_1$ - $\Sigma_5$ , respectively.

Following convention, we assume that a new systematic error with relative variance  $\tilde{\sigma}_{S,M}^2 = \tilde{\sigma}_{S,V}^2 + \tilde{\sigma}_{S,C}^2$  is generated after each measurement calibration. If there is only one calibration period, then we have a low-quality (one degree of freedom) estimate of  $\tilde{\sigma}_{S,M}^2 = \tilde{\sigma}_{S,V}^2 + \tilde{\sigma}_{S,C}^2$ . Regarding cleanout, a facility is typically (but not always) almost completely cleaned out, removing all inventory, approximately once per year for the annual physical inventory.

### 3.1. Avenhaus and Jaech (1981) Model ( $\Sigma_1$ )

Avenhaus and Jaech (1981) assume only random error components for  $I_t$  and  $T_t$ , inventory and transfer at the  $t$ th balance. There is no cleanout and no systematic measurement error. We have the  $t$ th balance (writing  $X_t$  for the MB at period  $t$ ),  $X_t = I_t + T_t - I_{t-1}$  where  $\sigma_{I,*}$  and  $\sigma_{T,*}$  are the absolute standard deviations of  $I_t$  and  $T_t$ , respectively. For all of our examples, we use the notation for the variance components from Avenhaus and Jaech (1981), which slightly modifies the notation in Section 2. We use the \* symbol as a dummy placeholder for the other four covariance scenarios to distinguish between systematic and random error components. Then  $\sigma_X^2 = 2\sigma_{I,*}^2 + \sigma_{T,*}^2$  for all  $t$  and  $Cov(X_{t-1}, X_t) = Cov(X_t, X_{t+1}) = -\sigma_{I,*}^2$ . Avenhaus and Jaech (1981) use values  $\sigma_{I,*} = 1$  and  $\sigma_{T,*} = 0.5$  based on 12 balance periods per year. If there are  $n$  balance periods per year, then  $\sigma_{T,*}^2 = (12/n)0.5^2$ . To be consistent with Avenhaus and Jaech (1981), we also use absolute variances

for the remaining four scenarios. In applications, the measured inventories, inputs, and outputs are used to convert between relative and absolute variances. Therefore, there is no loss of generality to simplify the presentation here and use absolute variances throughout. An additional simplification is to let the transfer term  $T$  represent the aggregate effect of input and output measurements. The remaining four scenarios have systematic measurement error.

### 3.2. Extended Model, No Cleanout, With Measurement Calibration ( $\Sigma_2$ )

The remaining four scenarios consider both random and systematic error components for  $I_t$  and  $T_t$ . Measurement calibration is assumed, for illustration purposes, to occur at the end of each year for both inventory and transfer measurements, but other recalibration schedules are certainly possible. We have the  $t$ th balance,  $X_t = I_t + T_t - I_{t-1}$ , where  $\sigma_{I,ran}$  and  $\sigma_{I,sys}$  are the random and systematic error standard deviations of  $I_t$ , while  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$  are the random and systematic error standard deviations of  $T_t$ , respectively. Then we have for the first year:

$$Var(I_t + T_t - I_{t-1}) = 2\sigma_{I,ran}^2 + \sigma_{T,ran}^2 + \sigma_{T,sys}^2,$$

$$Cov(I_t + T_t - I_{t-1}, I_{t+1} + T_{t+1} - I_t) = -\sigma_{I,ran}^2 + \sigma_{T,sys}^2,$$

and for  $j \geq 2$ ,

$$Cov(I_{t+j} + T_{t+j} - I_{t+j-1}, I_t + T_t - I_{t-1}) = \sigma_{T,sys}^2.$$

For multiple years, the covariance matrix is block diagonal, except for the last balance of one year and the first balance of the next year, where

$$Cov(I_n + T_n - I_{n-1}, I_{n+1} + T_{n+1} - I_n) = -\sigma_{I,ran}^2,$$

and except for the first balance of year one and the first balance of year two (and similarly for the first balance of year two and the first balance of year three, etc.), where

$$Cov(I_1 + T_1 - I_0, I_n + T_{n+1} - I_{n+1}) = -\sigma_{I,sys}^2,$$

The covariance between the last balance of one year and the first balance of the next year,  $Cov(I_n + T_n - I_{n-1}, I_{n+1} + T_{n+1} - I_n) = -\sigma_{I,ran}^2$ , is negative, because the ending inventory for one period is the beginning inventory for the next period, and there is no positive covariance arising from the systematic transfer measurement error because of the recalibration at the end of each year. The first balance of one year and the first balance of the next year are in separate calibration periods by assumptions, but  $Cov(I_1 + T_1 - I_0, I_{n+1} + T_{n+1} - I_n) = -\sigma_{I,sys}^2$ , is negative, because  $Cov(I_1 - I_0, I_{n+1} - I_n) = -Cov(I_1, I_n) = -\sigma_{I,sys}^2$ , with only  $I_1$  and  $I_n$  shar-

ing an inventory term systematic error. Notice also that  $\sigma_{I,sys}^2$  does not appear in the first-year expression because it cancels, but it does appear between successive years.

Appendix 1 gives the variance propagation results for the other three cases,  $\Sigma_3, \Sigma_4$ , and  $\Sigma_5$ .

## 4. BAYESIAN INFERENCE

In this article, we apply a Bayesian inferential approach. A Bayesian analysis combines prior information with observed data to produce a posterior distribution for the parameters  $\theta$  using Bayes' theorem:

$$\pi(\theta | \mathbf{y}) = \frac{L(\mathbf{y} | \theta) \pi(\theta)}{\int_{\Theta} L(\mathbf{y} | \theta) \pi(\theta) d\theta} \quad (1)$$

where  $\Theta$  denotes the range of values for the parameters  $\theta$ .  $\theta$  is the vector of  $\sigma_I$ ,  $\sigma_T$ ,  $\sigma_{I,sys}$ ,  $\sigma_{I,ran}$ ,  $\sigma_{T,sys}$  and  $\sigma_{T,ran}$ , where applicable. The likelihood function denoted by  $L(\mathbf{y} | \theta)$  describes the probability density function (pdf) of  $\mathbf{y}$  given the model parameters  $\theta$ , where  $\mathbf{y}$  denotes the vector of training data. The pdf of  $\mathbf{y}$  corresponds to

*MultivariateNormal*( $\mathbf{0}, \Sigma_{MB}$ ), so that  $L(\mathbf{y} | \theta)$  is this pdf, where  $\Sigma_{MB}$  is one of  $\Sigma_1$ - $\Sigma_5$ . The available information about  $\theta$  is summarized by the prior distribution  $\pi(\theta)$ . Bayes' theorem in Equation (1) demonstrates how the data and prior information are combined to obtain the posterior distribution of  $\theta$  denoted by  $\pi(\theta | \mathbf{y})$ .

In many applications, an analytical expression for the integral in Equation (1) does not exist. Instead, Markov chain Monte Carlo (MCMC) is used to simulate samples  $\{\theta^{(m)}; m = 1, \dots, M\}$  from the posterior distribution  $\pi(\theta | \mathbf{y})$ . See Casella and George (1992), Chib and Greenberg (1995) and Gelman et al. (2003) for discussions of popular MCMC algorithms.

Estimates of the parameters and their uncertainties are then based on the posterior samples. For each model parameter, empirical quantiles of its posterior samples, e.g., the 0.025 and 0.975 quantiles, are calculated to obtain central 95% credible intervals. The median or 0.5 quantiles of its posterior samples can be used as estimates. The width of the 95% credible intervals is a good measure of uncertainty, which in our context is the uncertainty on the standard deviations  $\sigma_{I,sys}, \sigma_{I,ran}, \sigma_{T,sys}$  and  $\sigma_{T,ran}$ .

## 5. SIMULATION STUDY

Here, we consider cases  $\Sigma_2$  -  $\Sigma_5$  with systematic measurement errors. Their  $\Sigma_{MB}$  depend on all or some of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ . We use lognormal priors on  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$  because these are positive quantities. Also, to allow for the prior to have the correct or an incorrect mean value, we consider two versions of the prior

distributions. The two prior versions, referred to as Prior 1 and Prior 2 are:

1.  $\sigma_{I,ran} \sim \text{Lognormal}(\log(\sigma_{I,ran,true}), 0.15^2)$ ,  
 $\sigma_{I,sys} \sim \text{Lognormal}(\log(\sigma_{I,sys,true}), 0.15^2)$ ,  
 $\sigma_{T,ran} \sim \text{Lognormal}(\log(\sigma_{T,ran,true}), 0.15^2)$ ,  
 $\sigma_{T,sys} \sim \text{Lognormal}(\log(\sigma_{T,sys,true}), 0.15^2)$ , i.e., the median of the prior is the true value and the prior is quite wide, because the variance of  $0.15^2$  is large.
2.  $\sigma_{I,ran} \sim \text{Lognormal}(\log(0.75\sigma_{I,ran,true}), 0.15^2)$ ,  
 $\sigma_{I,sys} \sim \text{Lognormal}(\log(0.75\sigma_{I,sys,true}), 0.15^2)$ ,  
 $\sigma_{T,ran} \sim \text{Lognormal}(\log(0.75\sigma_{T,ran,true}), 0.15^2)$ ,  
 $\sigma_{T,sys} \sim \text{Lognormal}(\log(0.75\sigma_{T,sys,true}), 0.15^2)$ , i.e., the median of the prior is 75% of the true value (biased low) and the prior is again quite wide.

We use  $m$  training data sets of two years of  $MB$  balances, where  $m = 2, 4, 10$ . We simulate  $m$  sets of data 500 times and  $I_0 = 0$  for all cases.

Also assume true  $\sigma_I = 1$  and  $\sigma_T = 0.1, 1$  with ratio  $\frac{\sigma_{ran}}{\sigma_{sys}} = 0.5$  and  $1.0$ . We consider  $n = 12$ , i.e., 12 balances per year. We use the Metropolis-Hastings algorithm to obtain draws from the joint posterior distribution of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$  (Chib and Greenberg, 1995) implemented in R (R Development Core Team, 2009). For each data, the Metropolis-Hastings algorithm is run for 3500 draws and the first 500 draws are discarded before summarizing the results.

To estimate the true values of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ , for each of the 500 simulated data sets, we use the posterior median of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ . These 500 median values can in turn be used to evaluate  $\Sigma_{MB}$ . The average over these 500 median values are viewed as average estimates of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$ ,  $\sigma_{T,sys}$ , and the elements of  $\Sigma_{MB}$ . For plotting purposes, for each data set, we divide the posterior medians of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$  by their true values to obtain relative medians. For each data set, we evaluate  $\Sigma_{MB}$  with the posterior medians of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ , element-wise divide by the corresponding non-zero elements of the true  $\Sigma_{MB}$ , and compute the median of these relative quantities. These quantities are then averaged over the 500 data sets. Figure 1 plots these for  $\Sigma_2$ , as explained further in Section 6.

Because the Prior 1 medians are the true values of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ , we expect values near 1. Because the Prior 2 medians are 75% of the true values of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ , we expect values less than 1 that increase as the number of training sets increases.

From the posterior distributions of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ , we divide the length of the 95% central credible intervals by that corresponding to their prior distributions. Taking the average over the 500 data sets provides an average relative length of the 95% central credible intervals. We expect values less than 1 that decrease as the number of training sets

increases. Figure 2 in Section 6 plots these types of summaries of the uncertainty in  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ .

## 6. SIMULATION STUDY RESULTS

### 6.1. $\Sigma_2$

Qualitatively, we expect the dashed curves to approach 1 in Figure 2, because eventually there is enough training data to overwhelm the wrong mean in Prior 2. Unfortunately, the approach to 1 is very slow; fortunately, the Prior 2 case is very pessimistic, but is included here for completeness.

The average relative posterior median results for Prior 2 are among the most interesting results in Figure 1. The best improvement is for  $\sigma_{I,ran}$ . From Figure 1c for  $\sigma_{T,ran}$ , there appears to be no improvement for (ratio,  $\sigma_T$ ) pairs (0.5,0.5) and (1,0.5), i.e., for the small  $\sigma_T$ . In Section 3.2,  $\sigma_T$  appears with other terms, making it hard to estimate when it is small. However, for the other two pairs, there is improvement. For  $\sigma_{I,sys}$  and  $\sigma_{T,sys}$ , there is only slight improvement. Figure 1e illustrates that there is little improvement for the elements of  $\Sigma_{MB}$  for Prior 1. There is improvement in the elements of  $\Sigma_{MB}$  for Prior 2.

Recall that Figure 2 plots the uncertainty in posterior distribution for  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ , ratioed to the uncertainty in the prior. Therefore, ideally the curves in Figure 2 should decrease toward zero as the training data size increases. Like Figure 1, Figure 2 illustrates the most improvement for  $\sigma_{I,ran}$ . There is no improvement for  $\sigma_{T,ran}$  for (ratio,  $\sigma_T$ ) pairs (0.5,0.5) and (1,0.5) as shown in Figure 2c.

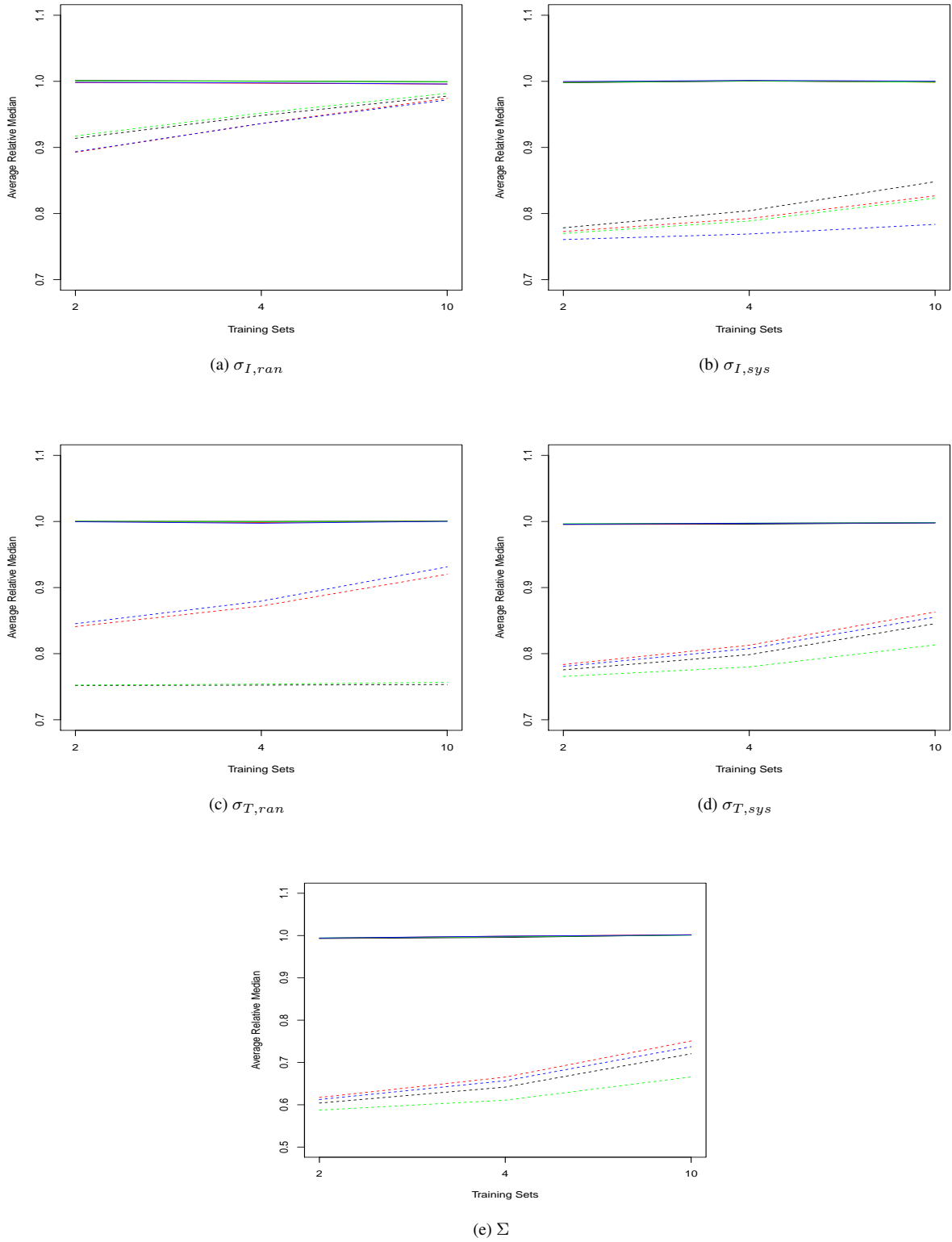
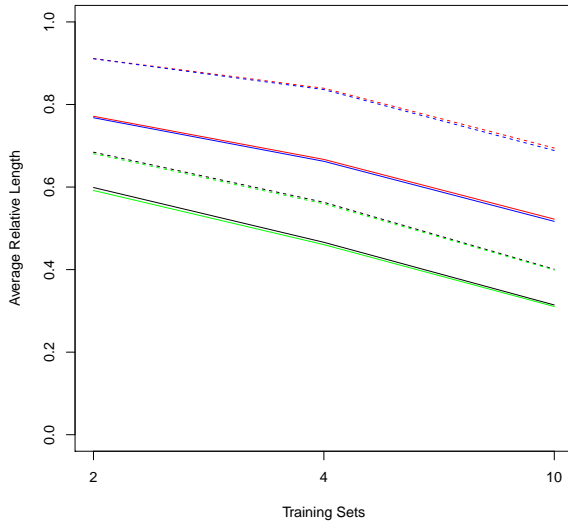
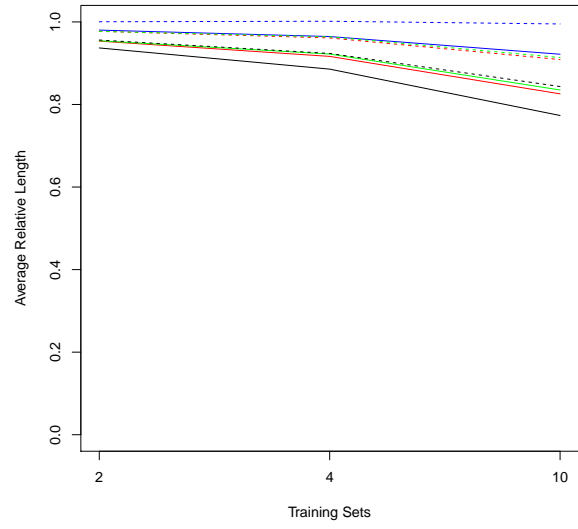


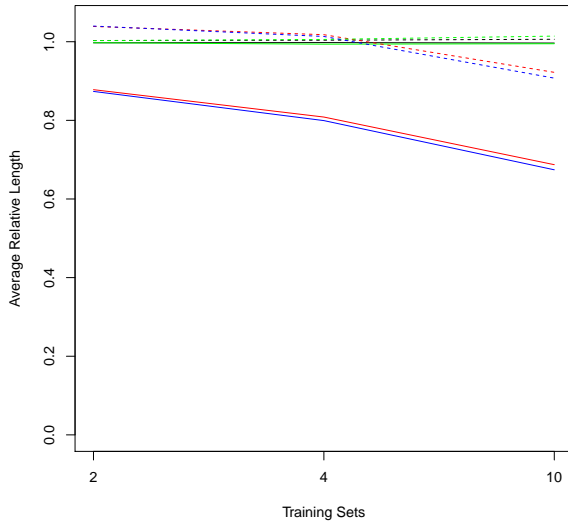
Figure 1. Average Relative Posterior Median for  $\Sigma_2$  (Colors black, red, green, blue correspond to (ratio,  $\sigma_T$ ) pairs (0.5,0.5), (0.5,1), (1,0.5), (1,1). Solid lines are for Prior 1. Dashed lines are for Prior 2.)



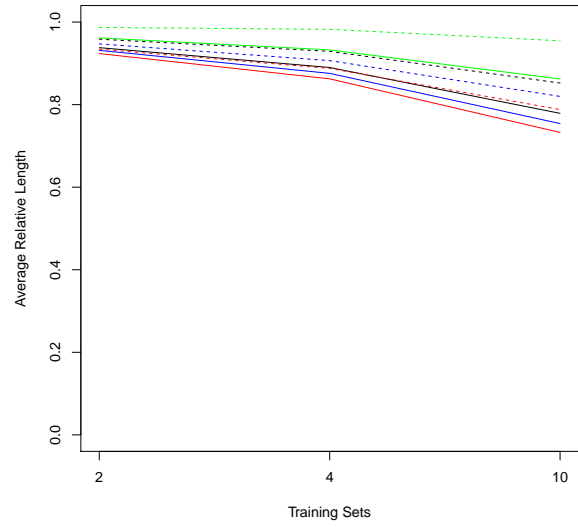
(a)  $\sigma_{I,ran}$



(b)  $\sigma_{I,sys}$



(c)  $\sigma_{T,ran}$



(d)  $\sigma_{T,sys}$

Figure 2. Average Relative Posterior Interval Length for  $\Sigma_2$  (Colors black, red, green, blue correspond to (ratio,  $\sigma_T$ ) pairs (0.5,0.5), (0.5,1), (1,0.5), (1,1). Solid lines are for Prior 1. Dashed lines are for Prior 2.)

## 6.2. $\Sigma_3$

The average relative posterior median results for Prior 2 are the most interesting results. The same type of figures just described for  $\Sigma_2$  are available in Burr and Hamada (2014c) for the  $\Sigma_3$ ,  $\Sigma_4$ , and  $\Sigma_5$  scenarios; we summarize here the key findings.

The best improvement is for  $\sigma_{I,ran}$ , while there is no improvement for  $\sigma_{I,sys}$ . This is not surprising because  $\sigma_{I,sys}$  does not appear in  $\Sigma_3$  as presented in Section 10.1, and so it is not important to estimate  $\sigma_{I,sys}$  well for our main purpose of improving the estimate of  $\Sigma_3$ .

There is no improvement for (ratio,  $\sigma_T$ ) pairs (0.5,0.5) and (1,0.5), i.e., for the small  $\sigma_T$ . In Section 3.2,  $\sigma_T$  appears with other terms, making it hard to estimate when it is small. However, for the other two pairs, there is improvement. For  $\sigma_{T,sys}$ , there is only slight improvement. and there is little improvement for the elements of  $\Sigma_{MB}$ .

## 6.3. $\Sigma_4$

The average relative posterior median results for Prior 2 are the most interesting findings. The best improvement is for  $\sigma_{I,ran}$ . There appears to be no improvement for (ratio,  $\sigma_T$ ) pairs (0.5,0.5) and (1,0.5), i.e., for the small  $\sigma_T$ . In the Appendix (Section 10.2),  $\sigma_T$  appears with other terms, so it is hard to estimate when it is small. However, for the other two pairs, there is improvement. For  $\sigma_{I,sys}$  and  $\sigma_{T,sys}$ , there is only slight improvement, and there is little improvement for the elements of  $\Sigma_{MB}$ .

## 6.4. $\Sigma_5$

Again the Prior 2 results are the most interesting. The best improvement is for  $\sigma_{I,ran}$ . For  $\sigma_{T,ran}$ , unlike that for the previous  $\Sigma_{MB}$ , there is improvement for (ratio,  $\sigma_T$ ) pairs (0.5,0.5) and (1,0.5). It is not clear from Section 10.3 in the Appendix why this addition to the variance propagation results in Section 10.2 makes a difference. For  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ , there is only slight improvement. There is some improvement for the elements of  $\Sigma_{MB}$ .

## 7. USING ONE LONG TRAINING SET

Here we consider the training set as one long data set rather than using blocks of one-year periods as in the previous option. Our purpose is to investigate which option leads to more improvement in the estimate of  $\Sigma_{MB}$ . As with the previous option, we use ratio  $\frac{\sigma_{ran}}{\sigma_{sys}} = 1$ ,  $\sigma_I = 1$ ,  $\sigma_T = (0.1, 1)$ ,  $n = 12$  balances, number of years training 4 and 8 as above, with Prior 2. From Section 10.2 in the Appendix, there is no difference for  $\Sigma_4$  between the one long training set and multiple segments covariance matrices, because they are block diagonal. Consequently, we do not consider  $\Sigma_4$  further. Note that for  $\Sigma_2$ ,  $\Sigma_3$  and  $\Sigma_5$ , the one long training set and multiple seg-

ments data will not be the same because their MB covariance matrices are different. Here we use average relative median and average relative interval length as previously considered. These results are based on 1000 data sets and 5000 posterior draws per data set.

## 7.1. $\Sigma_2$

From Section 3.2 with the one long training set, there are 1 and 3 additional covariance matrix entries over  $m = 2$  and 4 multiple two-year sets (i.e., between the last balance of the last year of the first set and the first balance of the first year of the next set) for  $\sigma_{I,ran}^2$  and the same for  $\sigma_{I,sys}^2$ . Consequently there is hardly any difference for the average relative posterior median results. The same holds for the posterior interval length results that are slightly smaller for the one long training set. These differences for  $\sigma_{I,ran}^2$  and  $\sigma_{I,sys}^2$  appear to be within “simulation error,” the error associated with doing a finite number of simulations. Similarly, apparent differences for  $\sigma_{T,ran}^2$  and  $\sigma_{T,sys}^2$  are easily attributed to “simulation error” associated with doing a finite number of simulations.

## 7.2. $\Sigma_3$

From Section 10.1 in the Appendix with the one long training set, there are 1 and 3 additional covariance matrix entries over  $m = 2$  and 4 multiple two-year sets (i.e., between the last balance of the last year of the first set and the first balance of the first year of the next set) for  $-\sigma_{I,ran}^2 + \sigma_{T,sys}^2$ , so there is little additional information for  $\sigma_{I,ran}^2$  besides it being coupled with  $\sigma_{T,sys}^2$ . However, for the remaining balances between the multiple data sets, their covariance matrix entries are  $\sigma_{T,sys}^2$ . We do not have an explanation for why one long period is worse for  $\sigma_{T,sys}^2$  and for  $\sigma_{T,ran}^2$ , especially for  $\sigma_T = 1$ ; however, the difference is small and possibly due mostly to using a finite number of simulations.

## 7.3. $\Sigma_5$

From Section 10.3 in the Appendix, there are additional covariance matrix entries between balances from different years that couple  $\sigma_{I,sys}^2$  and  $\sigma_{T,sys}^2$ . There are also additional covariance matrix entries between balances from different years for  $\sigma_{T,sys}^2$ . Again, using one long period is slightly worse for  $\sigma_{I,sys}^2$  and  $\sigma_{T,sys}^2$ .

## 8. SEQUENTIAL TESTING IN NMA

Quantitative evaluations of the loss detection probability for sequential tests of MB sequences with various loss scenarios have been presented (Avenhaus and Jaech, 1981; Beedgen, 1987; Jones, 1989; Speed and Culpin, 1986; Burr et al., 1995), but estimation error in  $\Sigma_{MB}$  was ignored. The most commonly-used sequential test in safeguards is Page’s cusum test applied to a transform of the MB sequence (known as



the SITMUF sequence, the standardized independently transformed MUF sequence, where MUF = MB is the international term “material unaccounted for”). Other sequential tests include the “power one” test, the cumulative MUF test, and the GeMUF test, which is similar to the Mahalanobis distance, but adapted to be sequential (Speed and Culpin, 1986; Beedgen, 1987;). For our purposes, it is sufficient to consider only Page’s cusum test applied to the SITMUF sequence. Our focus is to evaluate the impact of estimation error in  $\Sigma_{MB}$  (with or without estimation error reduction by using training data) on the performance of Page’s test applied to the SITMUF sequence.

Page’s statistic for SNM loss is calculated on balance period  $t$  as  $S_t = \max(0, S_{t-1} + Y_t - k)$ , where  $k$  is a control parameter and  $Y = AX$  is the SITMUF sequence computed from the MB sequence  $X = x_1, x_2, \dots, x_n$ . The matrix  $A = \Sigma_{MB}^{-1/2}$ , and in practice, the Cholesky decomposition is used to compute  $Y$ . Page’s test alarms at period  $t$  if  $S_t > h$  for some alarm threshold  $h$ .

The typical approach to apply Page’s test in NMA is to set  $k$  equal to approximately one half of a “significant quantity” (SQ), which for Pu is 8 kg, because this leads to a good detection probability for a loss of 8 kg or more (Page, 1955). Then, simulation is used to select  $h$  for  $k = 8/2 = 4$  to have the nominal false alarm probability (FAP) of 0.05 per year. Note that this approach assumes  $\hat{\Sigma}_{MB} = \Sigma_{MB}$ . If instead we allow for estimation error in  $\hat{\Sigma}_{MB}$ , we can estimate the range of FAPs that could result from estimation error in  $\hat{\Sigma}_{MB}$ . As an example, we assumed a 10 % relative standard deviation in each of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$  for scenario 2 with  $\Sigma_2$ . We found (by using  $10^6$  simulations, with each simulation having a different value of  $\hat{\Sigma}_{MB}$  according to the 10% relative standard deviation in each of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ ), that the lower 2.5% and upper 97.5% quantiles of the FAPs are 0.02 and 0.28, respectively, with an average FAP of 0.05, equal to the nominal FAP. Now if we multiply each of the 10% relative standard deviations by a factor of 0.8 (Figure 2 suggests that the factor 0.8 is reasonable for  $\sigma_{I,ran}$ , and is optimistic, depending on the prior and case, for the other standard deviations, but is used only for illustration here), then the 2.5% and upper 97.5% quantiles are 0.02 and 0.09, again with an average FAP of 0.05.

Similarly, the detection probability (DP) for a given SNM loss can be estimated using simulation, but estimation error in  $\hat{\Sigma}_{MB}$  is typically ignored. As with the FAP, we can use simulation to find the range of actual DPs that results from error in  $\hat{\Sigma}_{MB}$ . For example, we simulated the effect of SNM loss of 0.8 units (in units of the standard deviation) for 5 consecutive periods, again for case 2 scenario with  $\Sigma_2$ . Using simulation, we find that the true DP is 0.32 when the correct value of  $\Sigma_2$  is known. The 2.5% and upper 97.5% quantiles are 0.09 to 0.88, and the average DP is the same as the true

DP, 0.32. If the 10% relative standard deviations in  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$  are multiplied by a factor of 0.8, then the 2.5% and upper 97.5% quantiles are 0.29 and 0.34, respectively, again with an average of 0.32, which is the true DP.

If one uses the correct prior, our prior 1, then on average, the true FAP is the same as the nominal FAP. However, for any one facility, there could be a very large difference between the actual FAP and the nominal FAP (0.02 to 0.28 in our scenario 2 example, with a nominal FAP of 0.05 per year), and similarly for the DP. Current efforts to use PM data more quantitatively are confronted with the same challenge to try to make the actual FAP close in value to the nominal FAP.

## 9. DISCUSSION

This article demonstrated how training data can be used to estimate  $\Sigma_{MB}$ . Previous studies have ignored estimation error in  $\Sigma_{MB}$ . The simulation study demonstrated much improvement for the estimate of  $\sigma_{I,ran}$ . However, even with 20 years (20 calibration periods) of training data, there was only moderate or little improvement in the estimates of  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$ . We do not claim that these are general findings, although an argument based on degrees of freedom suggests that systematic error variances are more difficult than random error variances to estimate.

In recent efforts to apply PM, similar variance propagation equations as presented here apply (Smith et al., 2013). And, the PM is collected 10 times per hour, plus recalibrated frequently, so similar analyses are likely to be relevant for that type of PM data, although our treatment was for NMA data.

Quantitative PM evaluates the amount of training data needed in order for the actual FAP be close in value to the nominal FAP. Similarly, NMA should be challenged to illustrate that its actual FAP is close to the nominal FAP, particularly because the impact of estimation error in  $\Sigma_{MB}$  can get magnified when one inverts  $\Sigma_{MB}$  in order to compute the SITMUF. We have shown how training data can be used to improve the estimate of  $\Sigma_{MB}$ , but one conclusion is that metrology data should continue to be collected, as it provides the other option to improve the estimate of  $\Sigma_{MB}$ . For NMA data, 20 calibration periods (perhaps years) of training data is unrealistic, and was only examined for completeness. We did use diffuse prior distributions, however. For real processing plants, the measurement system should be well understood, so that the prior distributions are much tighter and this type of analysis could be repeated.

This article shows to what extent training MB data can be used to update the estimates of  $\sigma_{I,ran}$ ,  $\sigma_{I,sys}$ ,  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$  depending on the situation. Instead of training data consisting of multiple segments, we also tried one long data set. There were only slight differences between the multiple

segments and one long training set methods which we believe is due to using a finite number of simulations, based on 1000 data sets and 5000 MCMC posterior draws of the parameters per data set.

#### ACKNOWLEDGEMENTS

We acknowledge support from the National Nuclear Security Administration office of Nuclear Nonproliferation Research and Development (NA-22) and Nuclear Energy (NE) programs.

#### REFERENCES

- Aigner, H., et al. (2010). International Target Values 2010 for Measurement Uncertainties in Safeguarding Nuclear Materials, *Journal of Nuclear Materials Management* 30(2), available at [www.inmm.org](http://www.inmm.org).
- Avenhaus, R. and Jaech, J. (1981). On Subdividing Material Balances in Time and/or Space, *Journal of Nuclear Material Management*, 10, 24-33.
- Beedgen, R. (1987). Present and Future Aspects of PROSA - a Computer Program for Near Real Time Accountancy, *Proceedings of 28th Annual Meeting of the Institute of Nuclear Materials Management*.
- Burr, T., Hamada, M.S. (2014a). Smoothing and Time Series Modeling of Nuclear Material Accounting Data for Protracted Diversion Detection, to appear *Nuclear Science and Engineering*.
- Burr, T., Hamada, M.S., Ticknor, L., Weaver, B. (2014b). Model Selection and Change Detection for a Time-varying Mean in Process Monitoring, *Nuclear Instruments and Methods in Physics Research A* 751, 79-87.
- Burr, T., Hamada, M.S. (2014c). Bayesian Updating of the Material Balances Covariance Matrices with Training Data, Los Alamos National Laboratory report LA-UR-14-22154.
- Burr, T. and Hamada, M.S. (2013a). Revisiting Statistical Aspects of Nuclear Material Accounting. *Science and Technology of Nuclear Installations*, Article ID 961360, 15 pages, DOI:10.1155/2013/961360.
- Burr, T., Hamada, M.S., Howell, J., Skurikhin, M., Ticknor, L., Weaver, B. (2013b). Estimating Alarm Thresholds For Process Monitoring Data Under Different Assumptions About The Data Generating Mechanism, *Science and Technology of Nuclear Installations*, Volume 2013, Article ID 705878, 18 pages, DOI:10.1155/2013/705878.
- Burr, T., Hamada, M.S., Skurikhin, M., Weaver, B. (2012). Pattern Recognition Options To Combine Process Monitoring And Material Accounting Data In Nuclear Safeguards, *Statistics Research Letters* 1(1), 6-31.
- Burr, T., Coulter, C., Hakkila, E., Ai, H., Kadokura, I., Fujimaki, K. (1995). Statistical Methods for Detecting Loss of Materials Using Near-Real Time Accounting Data, *Proceedings of the 36th Annual Meeting of the Institute of Nuclear Materials Management*.
- Casella, G. and George E.I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46, 327-335.
- Chib, S. and Greenberg, E. (1995). Understanding the Metropolis Hastings Algorithm. *The American Statistician*, 49, 327-335.
- Downing, D.J., Pike, D.H., and Morrison, G.W. (1978). Analysis of MB data Using ARIMA models, *Nuclear Materials Management*, 7(4), 80-86.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis, Second Edition*. Boca Raton: Chapman and Hall.
- Goldman, A.S., Picard, R.R., and Shipley, J.P. (1992). Statistical Methods for Nuclear Materials Accounting: an Overview, *Technometrics*, 24(4), 267-281.
- Grznar, J., Booth, D.E., and Sebastian, P. (1997). The Use of Robust Smoothers in Nuclear Material Safeguards, *Journal of Chemical Information and Computer Sciences*, 37, 236-240.
- Hamburg, J.H., Booth, D.E., and Weinroth, G.J. (1996). A Neural Network Approach to the Detection of Nuclear Material Losses, *Journal of Chemical Information and Computer Sciences*, 36, 544-553.
- Howell, J., Bevan, G., Burr, T. (2013). Inspector Interfaces to Facilitate Subjective Reasoning about Quantities in Trends, *Computers and Chemical Engineering*, 48, 29-39.
- Jones, B. (1989). Near Real Time Materials Accountancy Using SITMUF and a Joint Pages Test: Improvement of the Test, *ESARDA Bulletin*, 16, 13-19.
- Page, E. (1955). A Test for a Change in a Parameter Occurring at an Unknown Point, *Biometrika* 42, 523-527.
- Prasad, S., Booth, D., Hu, M.Y., and Deligonul, S. (1995a). The Detection of Nuclear Materials Losses. *Decision Sciences*, 26, 265-281.
- Prasad, S., Booth, D., and Hu, M.Y. (1995b). Monitoring the Quality of a Chemical Production Process Using the Joint Estimation Method, *Journal of Chemical Information and Computer Sciences*, 35, 53-58.
- R Development Core Team (2009). R: a Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. (<http://www.R-project.org>).
- Smith, E., Lebrun, A., Labella, R. (2013). Potential Roles for Unattended Safeguards Instrumentation at Centrifuge Enrichment Plants, *Journal of Nuclear Materials Management* 32(1), 38-58.

Speed, T., Culpin D. (1986). The Role of Statistics in Nuclear Materials Accounting: Issues and Problems, *Journal of the Royal Statistical Society A*, 149(4), 281-313. 1986.

## 10. APPENDIX 1

Appendix 1 describes the assumptions for  $\Sigma_3$ ,  $\Sigma_4$ , and  $\Sigma_5$ .

### 10.1. Extended Model, No Cleanout, No Measurement Calibration ( $\Sigma_3$ )

We have the  $t$ th balance,  $X_t = I_t + T_t - I_{t-1}$  where  $\sigma_{I,ran}$  and  $\sigma_{I,sys}$  are the random and systematic error standard deviations of  $I_t$  and  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$  are the random and systematic error standard deviations of  $T_t$ . Then we have:

$$Var(I_t + T_t - I_{t-1}) = 2\sigma_{I,ran}^2 + \sigma_{T,ran}^2 + \sigma_{T,sys}^2,$$

$$Cov(I_t + T_t - I_{t-1}, I_{t+1} + T_{t+1} - I_t) = -\sigma_{I,ran}^2 + \sigma_{T,sys}^2,$$

and for  $j \geq 2$ ,

$$Cov(I_{t+j} + T_{t+j} - I_{t+j-1}, I_t + T_t - I_{t-1}) = \sigma_{T,sys}^2.$$

This holds for multiple years.

### 10.2. Extended Model, Cleanout, Measurement Calibration ( $\Sigma_4$ )

We have the  $t$ th balance,  $X_t = I_t + T_t - I_{t-1}$  where  $\sigma_{I,ran}$  and  $\sigma_{I,sys}$  are the random and systematic error standard deviations of  $I_t$  and  $\sigma_{T,ran}$  and  $\sigma_{T,sys}$  are the random and systematic error standard deviations of  $T_t$ . Because of cleanout,  $I_0$  and  $I_n$  are zero with no associated measurement error. Then we have for the first year:

$$Var(I_1 + T_1) = \sigma_{I,ran}^2 + \sigma_{I,sys}^2 + \sigma_{T,ran}^2 + \sigma_{T,sys}^2$$

$$Var(T_n - I_{n-1}) = \sigma_{I,ran}^2 + \sigma_{I,sys}^2 + \sigma_{T,ran}^2 + \sigma_{T,sys}^2$$

For  $2 \leq t \leq n-1$ ,

$$Var(I_t + T_t - I_{t-1}) = 2\sigma_{I,ran}^2 + \sigma_{T,ran}^2 + \sigma_{T,sys}^2.$$

$$Cov(I_1 + T_1, I_2 + T_2 - I_1) = -\sigma_{I,ran}^2 + \sigma_{T,sys}^2$$

For  $3 \leq t \leq n-1$ ,

$$Cov(I_1 + T_1, I_t + T_t - I_{t-1}) = \sigma_{T,sys}^2.$$

$$Cov(I_1 + T_1, T_n - I_{n-1}) = -\sigma_{I,sys}^2 + \sigma_{T,sys}^2$$

$$Cov(I_{n-1} + T_{n-1} - I_{n-2}, T_n - I_{n-1}) = -\sigma_{I,ran}^2 + \sigma_{T,sys}^2$$

For  $2 \leq t \leq n-2$ ,

$$Cov(I_t + T_t - I_{t-1}, T_n - I_{n-1}) = \sigma_{T,sys}^2.$$

For  $2 \leq t \leq n-2$ ,

$$Cov(I_t + T_t - I_{t-1}, I_{t+1} + T_{t+1} - I_t) = -\sigma_{I,ran}^2 + \sigma_{T,sys}^2.$$

For  $t \geq 2, j \geq 2$ , and  $t+j \leq n-1$ ,

$$Cov(I_t + T_t - I_{t-1}, I_{t+j} + T_{t+j} - I_{t+j-1}) = \sigma_{T,sys}^2.$$

For multiple years, the covariance matrix is block diagonal.

### 10.3. Extended Model, Cleanout, No Measurement Calibration ( $\Sigma_5$ )

We have the same covariances for balances within the same year as given in the preceding scenario. Below is what has to be added for years 1 and 2 for all pairs of years.

$$Cov(I_1 + T_1, I_{n+1} + T_{n+1}) = \sigma_{I,sys}^2 + \sigma_{T,sys}^2$$

$$Cov(T_n - I_{n-1}, I_{n+1} + T_{n+1}) = -\sigma_{I,sys}^2 + \sigma_{T,sys}^2$$

For  $n+2 \leq k \leq 2n-1$ ,

$$Cov(I_1 + T_1, I_k + T_k - I_{k-1}) = \sigma_{T,sys}^2.$$

$$Cov(I_1 + T_1, T_{2n} - I_{2n-1}) = -\sigma_{I,sys}^2 + \sigma_{T,sys}^2$$

For  $2 \leq j \leq n-1$ ,

$$Cov(I_j + T_j - I_{j-1}, T_{2n} - I_{2n-1}) = \sigma_{T,sys}^2.$$

$$Cov(T_n - I_{n-1}, T_{2n} - I_{2n-1}) = \sigma_{I,sys}^2 + \sigma_{T,sys}^2$$

For  $2 \leq j \leq n-1$  and  $n+2 \leq k \leq 2n-1$ ,

$$Cov(I_j + T_j - I_{j-1}, I_k + T_k - I_{k-1}) = \sigma_{T,sys}^2.$$

For  $2 \leq j \leq n - 1$ ,

$$\text{Cov}(I_j + T_j - I_{j-1}, I_{n+1} + T_{n+1}) = \sigma_{T,sys}^2.$$

For  $n + 2 \leq j \leq 2n - 1$ ,

$$\text{Cov}(T_n - I_{n-1}, I_j + T_j - I_{j-1}) = \sigma_{T,sys}^2.$$