Evaluating Safeguards Statistical Assumptions via Stochastic Simulation

C. Gazze[1], C. Norman[1], R. Binner[1], S. Aghara[2], L. Annadevula[2], L. Joyce[2], K. Jarman[3], J. Gomera[4], K. Bachner[4]

[1]International Atomic Energy Agency (IAEA), Department of Safeguards, Vienna, Austria
[2]University of Massachusetts Lowell, Lowell, MA, USA
[3]Pacific Northwest National Laboratory, Richland, WA, USA
[4]Brookhaven National Laboratory, Upton, NY, USA

Abstract

The authors built and tested a stochastic simulation to estimate achieved detection probabilities (DPs) on a stratum basis, over a tailorable range of diverted amounts from 0 to 2 significant quantities (SQ), using typical International Atomic Energy Agency (IAEA) inspection data: i.e., SQ in stratum, number of items, number of gross/partial/bias defect measurements conducted, and realistic relative standard deviation (RSD) values for typical IAEA verification measurements. For bulk strata, the model calculates achieved DP at 0.01 SQ diversion increments; for item strata, the model calculates DP using the smallest realistic diversion increment (e.g., a plate, pin, or coupon). After successfully benchmarking against IAEA deterministic models, the simulation was used to test the sensitivity of DP to certain standard assumptions and selected input parameters. First, the equal defect assumption was tested; the results suggest significant complexity in the effectiveness of partial defect measurements. Next, the authors explored the sensitivity of DP to the assumed RSD of attribute tests. Then, the authors compared non-normal models for instrument performance (e.g., logistic, step, or arbitrary functions) to the typical results from a normal distribution (characterized by RSD). This last comparison was supplemented with experimentally derived performance data for an HM-5 gamma spectrometer. The HM-5 was used to make enrichment measurements on both LEU and HEU MTR fuel elements as plates were removed, and the results fit with logistic and step curves and applied in the simulation. These stochastic DP results were compared to DP estimates from a deterministic model assuming a normal curve and typical RSD, yielding insights that could improve effectiveness in the field. These early results illustrate the potential of stochastic models to better understand achieved DP and to improve safeguards effectiveness.

Introduction

To verify nuclear material inventories in a facility, IAEA inspectors first group the items into different strata and then test (measure) randomly selected items in each stratum to look for differences (*defects*) between operator declarations and measurement results. The IAEA employs well known statistical sampling methods [1, 2, 3] to achieve a desired probability for detecting the diversion of a *significant quantity* (SQ)[1] of a given *material type*[2] from each stratum. This detection probability (DP) is equal to the probability that one or more defected items were selected in the sample (P_s) multiplied by the probability the test identifies a selected item as defected (P_i).

This statistical sampling approach can provide a prescribed number of measurements of different methods (a sampling plan) as well as the corresponding expected DPs for each stratum. Under very

limited circumstances, these stratum level DPs are equivalent to the detection probability for a 1 SQ diversion of that material type at the facility level (that is, equivalent to the probability of detecting the diversion of sub-SQ amounts from multiple strata that total 1 SQ at the facility level). But more generally this is not the case.

As IAEA safeguards evolve to focus more on the State level, there is a need to estimate achieved[3] DP over aggregated inventories and against more sophisticated diversion strategies that involve multiple strata. The goal of this research effort has been to examine the utility of stochastic methods for estimating achieved DP in such cases.[4] The development and benchmarking of a stochastic model to treat this problem are detailed in Refs [4, 5, 6]. This paper will summarize interesting early results from this stochastic model.

Background

The stratification process involves grouping nuclear material by common characteristics like material type, physical form (e.g., item vs bulk), and relative mass. Some of these groupings are related to the measurements to be made (inside a given stratum, sampled items must be measured with the same methods), and some are related to the statistical methods applied (e.g., the assumptions underlying these statistical methods break down if items with very different masses are grouped together).

Selected items from each stratum are subjected to *gross defect* measurements, and sometimes also to *partial* or *bias defect* measurements. A gross defect measurement tests for the diversion of all the nuclear material in an item. It is typically a qualitative, "yes/no" test (an *attribute test*) such as looking for the 186 keV gamma peak from U-235 or for the collimated Cherenkov light from spent fuel assemblies. A partial defect test is a quantitative measurement to determine if there is substantially less material in an item than declared. It is typically a non-destructive assay (NDA) such as the neutron coincidence counting performed on fresh LWR fuel assemblies. A bias defect test looks for the protracted diversion of small amounts of nuclear material that would go unnoticed in a typical NDA measurement. This has traditionally implied sampling for destructive analysis (DA) but could also include, for example, the use of gamma tomography to detect the diversion of a single pin from a spent fuel assembly.

Inspectors design sampling plans (how many randomly selected samples from each stratum to test) using the list of inventory items provided by an operator and estimates of the relative measurement uncertainties of the various measurements they will apply. These estimates are provided as relative standard deviations (RSDs) for each measurement method assuming a normal distribution.[5] For qualitative (gross defect) measurements, a 15% RSD has historically been assumed. Selected samples are subjected to gross, partial, or bias defects using a statistical approach that generally attempts to minimize the number of more intensive (bias and partial) measurements in favor of simpler gross defect measurements, while still delivering the desired overall DP in the stratum.

Summary of the Stochastic Model and Test Data

Development and benchmarking of the stochastic model are detailed in Refs [4, 5, 6]. In short, the model takes as input a set of facility data that specify (by stratum): the number of items; the mass of nuclear material in SQ; the number of gross, partial, and bias defect measurements performed during the inspection; and the RSDs of the applied measurements. Mock facility data were created for an

enrichment plant, a fuel fabrication plant, a light water reactor (LWR), and a research reactor (MTR). This data was derived from actual facilities, rounded, and then scaled to representative sizes. The measurement RSDs were also derived from real data but rounded and adjusted to be realistic but non-specific.

Using a Monte Carlo approach, the stochastic model repeatedly simulates the random selection and measurement of items during an inspection to determine the probability of detecting a given diversion. At first, the measurements were simulated using an RSD-based error model, but this was later generalized to handle any kind of error model. DPs were calculated for diversions ranging in size from 0 to 2 SQ at (nominally) 0.01 SQ increments. This granularity was incorporated to address the combination of sub-SQ diversions from multiple strata in later calculations. Presently, the stochastic model considers two diversion strategies: Case 1) equal amounts are diverted from every item in the stratum (to minimize identification probability), and Case 2) diversions are concentrated in as few items as possible (to minimize selection probability).[6] In Case 2, one item is partially diverted while the rest are either fully diverted or not tampered.

For comparison, IAEA statistical methods assume the defects across a set of tampered items are equal in mass (the "equal defect assumption"), and then calculate the DP for a 1 SQ diversion spread across a varying number of items from 1 to N (where there are N items in the stratum). The stochastic model was adapted to duplicate this scenario as well, which is a type of Case 1 hybrid. In fact, the stochastic model is able to model any arbitrary set of defects (varying both the number of tampered items and the amounts taken from each item), meaning it could be used to test the sensitivity of the IAEA model to the equal defect assumption.

The stochastic model outputs DP curves vs. SQ diverted (and, if desired, also vs. items tampered) for each stratum in a facility. The model then attempts to combine these curves using various algorithms to determine a minimum DP curve across all strata in a facility. Stratum level results were benchmarked against a deterministic model, the results of IAEA models, and other published data [5, 6] to ensure the stochastic model was functioning correctly. Aggregated results were also benchmarked against a published case employing a deterministic solution. [6]

Additional functionality was later added to more accurately model the attribute (yes/no) measurements typically used for gross defect tests (e.g., like detecting a U-235 gamma peak from a coupon or the Cherenkov glow of spent fuel.). To do this, step and logistic curves were substituted for RSD-based models of uncertainty. In theory, the stochastic approach can accept any arbitrary model for the identification probability of an instrument. To this point, the authors made enrichment measurements on both HEU and LEU MTR fuel assemblies as plates were removed one-by-one. These empirical results were then fit to a logistic or step curve and used in the stochastic model to represent the identification probability for these specific measurements. Later in this paper, the results of these measurements are compared to the results from standard gross defect attribute tests with an assumed 15% RSD.

Selected Results for an Enrichment Plant

Figure 1a shows the achieved detection probabilities for a stratum of 26 LEU product cylinders at an enrichment plant for both diversion strategies (Cases 1 and 2). The sampling plan was designed to achieve 20% DP against a 1 SQ diversion of LEU from this stratum, which it obviously does for both

diversion strategies. However, the sampling plan is less effective when diversions are concentrated in fewer cylinders (that is, minimizing the selection probability) than when a small amount is taken from all cylinders. The result highlights the effectiveness of—and the need for—the DA samples (bias defect measurements), which are almost certain to detect defects exceeding a small threshold, in this case, about 1 - 2% of the cylinder content.

Figure 1b shows the result for a stratum of 300 DU tails cylinders in the same facility. In this case, however, the sampling strategy is more effective when the diversion is concentrated in fewer cylinders. Spreading the diversion across all 300 cylinders creates defects so small that even the DA measurement is not sensitive enough to detect the diversion at the desired level. (It could be that a bias would be discernable across all measurements; the added DP from material balance evaluation was not considered in this simulation.)

Finally, Figures 2a and 2b show the Case 1 hybrid where the equal diversion assumption is applied but the number of defected items is varied. The resulting DP surfaces show an interesting feature—a trough of lower DP. The authors had originally thought that Cases 1 and 2 might bound the solution space, but the presence of this trough indicates that the tradeoff between selection and identification probabilities is more complex that initially believed. One hypothesis for this trough is that it reflects the impact of partial defect measurements (which combine features of gross and bias measurements). In any case, the result warrants further study, and stochastic simulation has proven to be a very useful tool for such explorations.

Testing Statistical Assumptions

As just illustrated, the stochastic model provides a highly flexible tool for exploring the sensitivity of DP to different sampling plans and measurement techniques. It can also be used to test the sensitivity of statistical models to their underlying assumptions. The IAEA statistical model often assumes a 15% RSD for gross defect measurements. Figure 3 shows the sensitivity of DP to a range of assumed RSD values from 10% to 30% for the gross defect verification of spent fuel at an LWR (Case 2). For this simulation each spent fuel assembly holds 0.6 SQ of plutonium (or about 0.002 SQ per pin), pins are diverted one at a time, and defects are concentrated in as few assemblies as possible. False alarms are removed (resolved) so that they do not artificially inflate the DP. We see a quantization effect in the results as each additional assembly is defected. Each diverted assembly adds about 10% DP, but where this shift occurs varies significantly with the assumed RSD. Note that while the size of this effect in this LWR stratum was significant (a 10% step over a 0.4 SQ range), in other cases the effect was smaller (e.g., for DU cylinders at an enrichment plant it was only 2%).

As a result, it would seem important to understand the sensitivity of DP to the assumed gross defect RSD on a case-by-case basis. Alternatively, a more accurate model could be adopted for such qualitative gross defect measurements. Another advantage of the stochastic approach is the ease by which different measurement models can be substituted. The stochastic model used here was ultimately generalized to take as input any arbitrary identification probability curve. In particular, step and logistic curves were eventually substituted for the normally distributed error model, and an MTR case was even run using an experimentally derived, instrument-specific random error model (see below).

MTR case including a non-normal error model

At research reactors, gross defect measurements are a commonly applied measure for verifying fresh MTR fuel assemblies. Inspectors typically use a handheld gamma spectrometer (an "HM-5") to confirm the presence of the 186 keV U-235 peak in a selected item. However, the same instrument, properly calibrated, could also be used to perform an enrichment measurement on the assembly.

For this simulation, the authors created two experimentally derived identification probability (P_i) curves: one for an attribute (yes/no) test for the presence of a 186 keV peak, and one for an enrichment measurement. For the attribute test, it was confirmed experimentally that the HM-5 used in attribute mode could easily detect a U-235 peak even when a single MTR plate was placed behind a dummy assembly. This experiment was used to design a step shaped identification probability curve where the presence of one or more MTR plates provided a positive result and the absence of all plates detected the defect.

For the enrichment measurement, the empirical P_i curve was built as follows: The HM-5 enrichment measurement was calibrated by measuring a normal (18 plate) HEU MTR assembly with a nominal enrichment of 93%. This measurement yielded an estimated random error of approximately 1%. Plates were then removed one at a time and the enrichment measurement repeated. Figure 4a shows the collected data with the 3σ cutoff used as an alarm threshold (horizontal red line). It can be seen that, as plates are removed, the altered geometry and the deficit of uranium seen by the HM-5 detector causes an apparent drop in the measured enrichment. This data was fit by a logistic-shaped identification probability curve (Figure 4b).

The authors then used these two experimentally derived curves (logistic and step) along with the standard 15% RSD error model to simulate the results of diversion from an 80-element fresh fuel stratum totaling 0.8 SQ of HEU. Specifically, the stochastic model simulated 1) an attribute (yes/no) measurement for the presence or absence of U-235 using a Gaussian error model with an RSD of 15% (the default way this measurement is planned and analyzed at IAEA), 2) the same attribute measurement using the experimentally validated step function, and 3) an enrichment measurement using the experimentally derived logistic function.

Figure 5 plots the results for two diversion strategies: a) diverting only full assemblies, and b) diverting plates but spreading the defects over as many assemblies as possible. In the case of full assembly diversion, all three methods yield identical DP curves, because DP only depends on the selection probability (all three identification probabilities are the same for a fully diverted item). However, when the diversion is spread over all the assemblies by removing individual plates the three simulations yield very different results. The 15% RSD assumption predicts a DP of about 50% once the defect exceeds about 0.33 SQ, whereas the step curve predicts that the DP remains effectively nil until the diversion exceeds 0.7 SQ (almost all of the plates removed). This implies that the Gaussian error model substantially overestimates the achieved DP for this verification and is not a good model on which to base sample planning. A more realistic step curve should be used instead.

The enrichment measurement achieves a much higher DP for much smaller defects (already 90% by the time the diversion exceeds 0.05 SQ). An attribute measurement typically takes about 10 seconds to complete while the enrichment measurement requires 180 seconds. However, this time difference is not significant when only a few measurements are needed during a one-day inspection (which is

typically the case). The implication is that enrichment measurements would be far more effective (deliver much higher DP) than attribute measurements for a similar investment of time, but also that better modelling is required to deliver accurate estimates of expected DP when creating sampling plans or selecting instruments and measurements to apply. Stochastic simulation could therefore contribute much to the planning and implementation of safeguards by allowing realistic modelling of verification measurements.

Aggregation of DP to Facility and State Levels

As mentioned at the beginning of this paper, the main purpose for investigating stochastic models was to see if they could be used to aggregate achieved DP in individual strata to the facility and State levels. Preliminary results are available with respect to this challenge [7, 8], but this work is more generally still in progress.

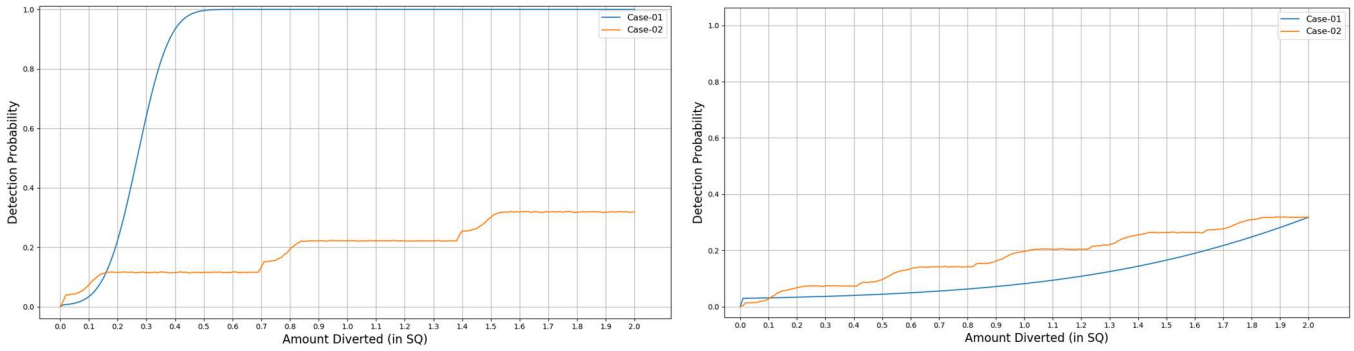
A number of algorithms have been tested to date in an attempt to produce accurate results using reasonable computing time. Figures 6 and 7 show the results of a partitions method that computes all possible combinations at each step (0.01 SQ) and then selects the lowest DP. This method is 100% accurate but computationally intensive. For the enrichment case a result was calculated in 2 minutes, but for a more complex set of strata and measurements at a fuel fabrication facility it took almost 30 minutes and only yielded results up to 0.5 SQ. In some cases, the model exceeded the available computational capacity.

The problem is tricky and some of the algorithms tried by the authors failed to yield the lowest DP (using the partitions method as a check). Initial test of a Greedy algorithm indicate that it works well. Figure 8 provides an example result. The authors continue to test other algorithms that can solve this minimization problem.

Conclusion

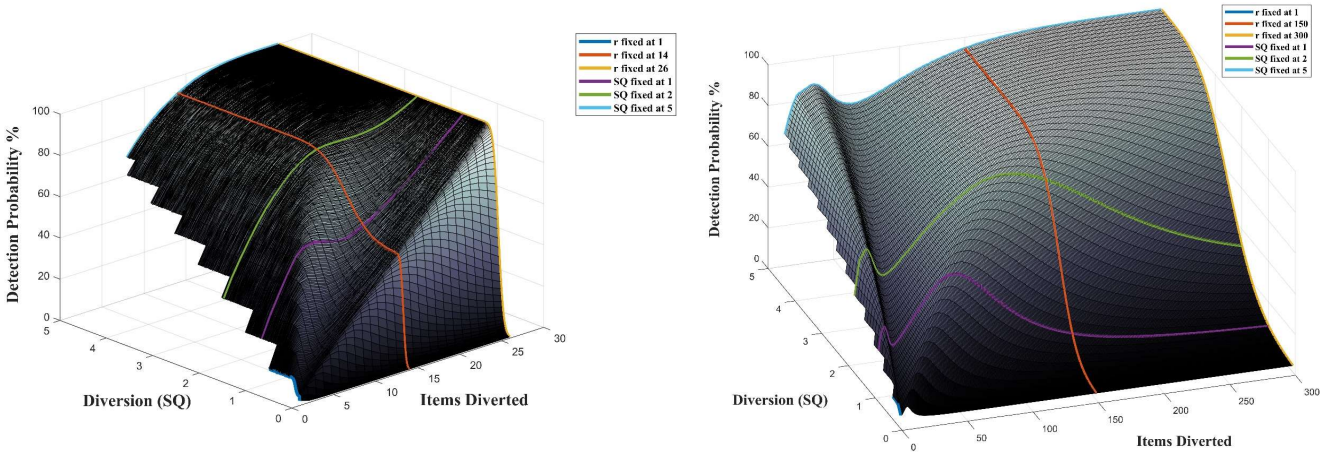
A stochastic simulation was built and tested to estimate achieved detection probabilities (DPs) on a stratum basis for typical IAEA verification measurements. While the main purpose for investigating stochastic models was to see if they could be used to aggregate achieved detection probabilities in individual strata to the facility and State levels, stochastic simulation has also proven useful for testing statistical assumptions and evaluating the effectiveness of selected verification measures. In particular, the authors confirmed that the tradeoff of identification and selection probabilities in certain diversion strategies can be quite complex, possibly due to the impact of the partial defect test. This result implies that the consequences of the equal diversion assumption warrant further study. Also, stochastic simulation was able to shed light on the sensitivity of DP to the error models and assumed RSD for gross defect tests. These results indicate that better modelling is required to deliver more accurate estimates of expected DP when creating sampling plans or selecting instruments and measurements to apply in the field.

The overall conclusion from our initial results is that stochastic simulation is a very promising tool that could contribute greatly to the calculation of achieved DP and the understanding of how different factors influence achieved detection probabilities.



(a) *(b)*

Figure 1: *Achieved detection probabilities at an example enrichment plant for (a) an LEU product stratum, and (b) a DU tails stratum. Results shown for both diversion strategies: minimizing identification probability (Case 1) and minimizing selection probability (Case 2).*



(a) *(b)*

Figure 2: *Case 1 hybrid diversion strategy showing achieved detection probabilities when equal amounts are taken from a varying number of items at an example enrichment plant: (a) LEU product stratum and (b) DU tails stratum. Compare to Figure 1.*

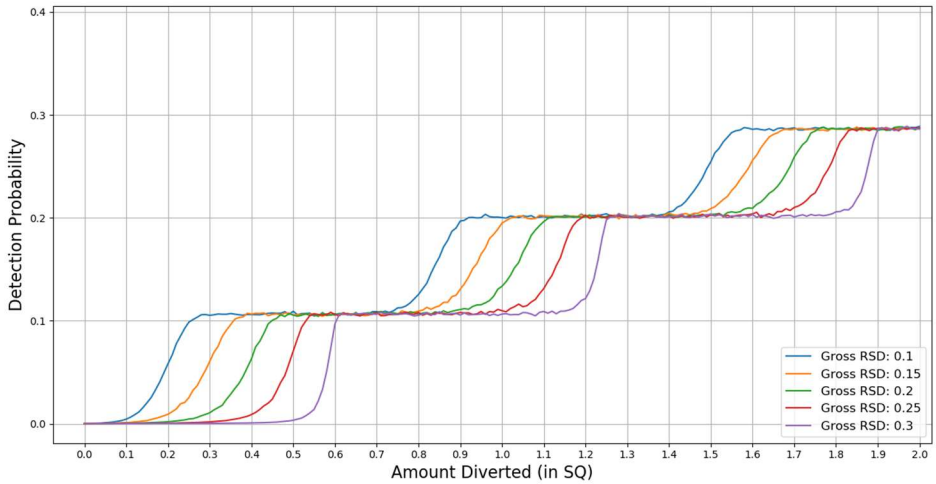
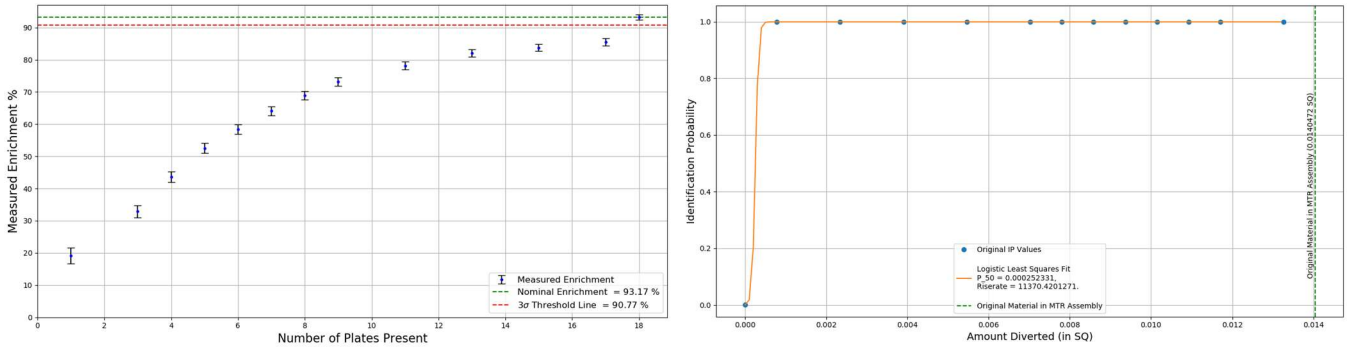


Figure 3: *Sensitivity of achieved detection probability to a range of assumed RSD values from 10% to 30% for gross defect verification of spent fuel at an LWR (Case 2) after the removal (resolution) of false alarms.*



(a) *(b)*

Figure 4: *(a) Experimental enrichment measurements of a fresh MTR fuel element (nominal enrichment = 93%) using an IAEA "HM-5" detector as plates are removed. Horizonal red line shows the 3σ alarm level under which a defect is considered detected. (b) Fit of MTR data to a logistic P_i curve.*

(a)

(b)

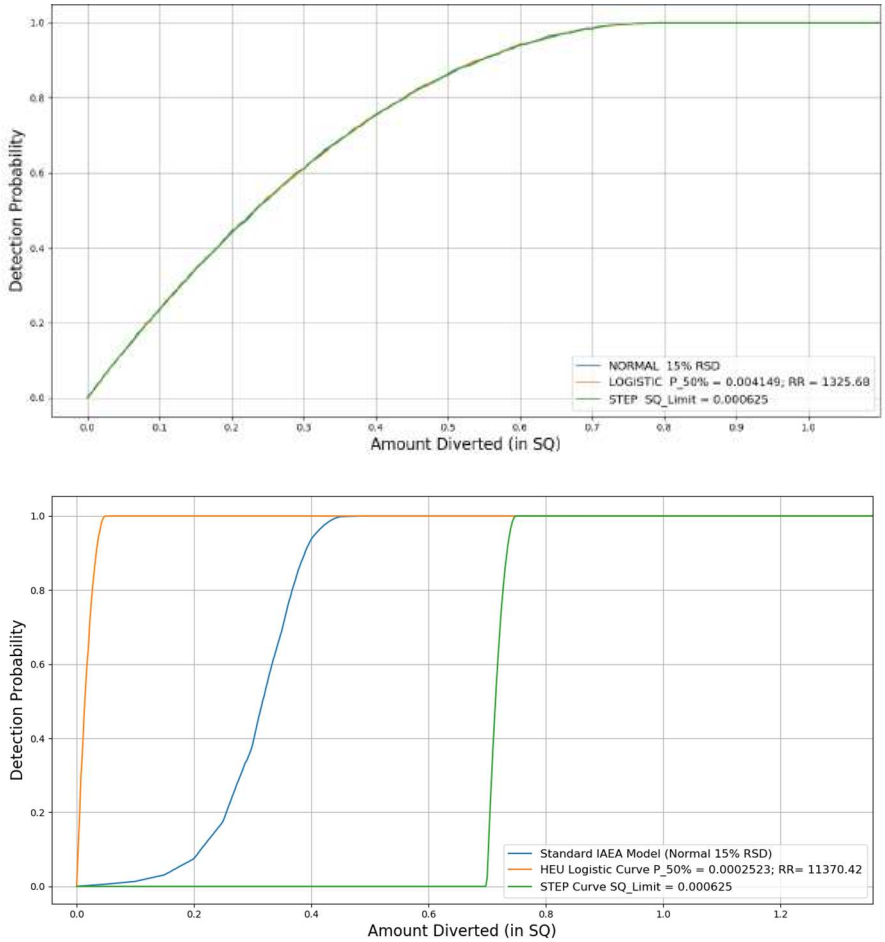


Figure 5: *Achieved detection probabilities for a fresh HEU MTR fuel stratum: (a) minimizing selection probability (Case 2) and (b) minimizing identification probability (Case 1). This figure compares results of attribute measurements with different error models (assumed 15% RSD vs experimentally determined step curve) against enrichment measurements (with an experimentally determined logistic error model).*

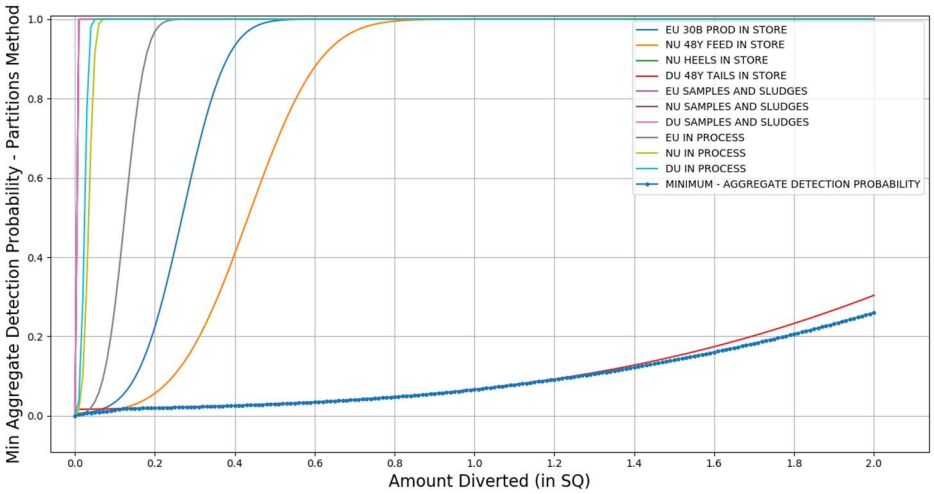


Figure 6: *Minimum aggregate detection probability estimated with the brute force (partitions) method for an example enrichment facility. The diversion strategy was to minimize identification probability (Case 1). Computation time was approximately 2 minutes.*

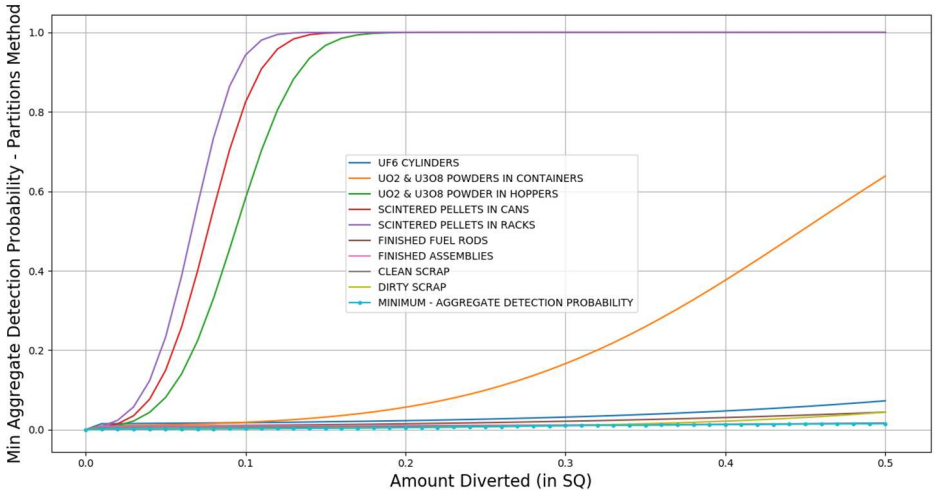


Figure 7: *Minimum aggregate detection probability estimated with the brute force (partitions) method for an example fuel fabrication facility with nine strata. The diversion strategy was to minimize identification probability (Case 1). Computation time was approximately 30 minutes.*

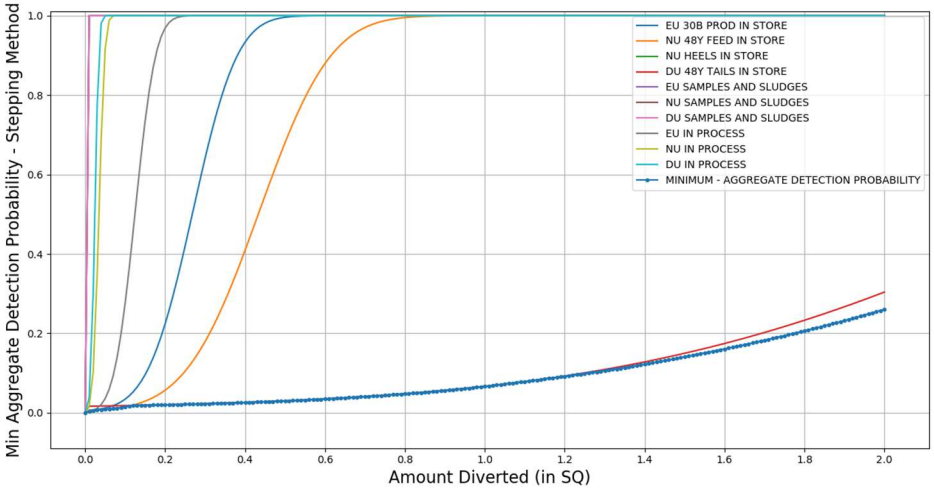


Figure 8: *Minimum aggregate detection probability estimated with steeping method or Greedy algorithm for the same enrichment facility as shown in Figure 6. The diversion strategy was to minimize identification probability (Case 1). Computation time was approximately 5 seconds.*

Notes:

[1] By definition, 1 SQ equals 8 kg of plutonium or U-233, 25 kg of U-235 in HEU, 75 kg of U-235 in LEU, 10 MT of natural uranium, 20 MT of depleted uranium, or 20 MT of thorium.

[2] Material types include plutonium, DU, NU, LEU, HEU, and thorium in either irradiated or non-irradiated forms.

[3] "Achieved" is used here to indicate post-inspection rather than a priori. In a statistical sense these are estimated detection probabilities.

[4] To date this project has limited its scope to considering the achieved detection probability from a given set of measurements. Another interesting problem is to optimize the required sampling plan to achieve a desired DP at the facility or State level. This more complex problem has not yet been addressed.

[5] RSDs are provided by IAEA statisticians and are generally derived from past performance data for the given measurement or for similar measurements or (alternatively) based on International Target Values (ITVs).

[6] It may be thought that these two strategies bound the problem. However, the relationship between these two can be more complex. Results indicating such a case will be presented later in the paper.

References:

1. Jaech, J. L., and M. Russell. *Algorithms to calculate sample sizes for inspection sampling plans*. No. I1. AEA-STR--261 (REV. 0). International Atomic Energy Agency, 1990.
2. IAEA - International Atomic Energy Agency. *Statistical Concepts and Techniques for IAEA Safeguards*, IAEA SG-PR-2016 Rev. 5, 1998.
3. T. Krieger, *et al.*, *Statistical Methods for Verification Sampling Plans*, IAEA STR-381 draft, 2017.
4. L. Annadevula, *et al.*, "Review of Methods to Aggregate Diversion Detection Probabilities Across Multiple Material Strata," submitted concurrently to *Proc. of the INMM & ESARDA Joint Annual Meeting*, 2021
5. L. Annadevula, *et al.*, "Statistical Analysis of Convergence and Error Propagation in Stochastic Model for Safeguards Inspection," submitted concurrently to *Proc. of the INMM & ESARDA Joint Annual Meeting*, 2021.
6. Joyce, L., *et al.*, "Stochastic Model Simulation for Evaluation of Spent Fuel Pond Inventory Verification Sampling Plans," submitted concurrently to *Proc. of the INMM & ESARDA Joint Annual Meeting*, 2021.
7. Krieger, T., *et al.*, "Multi-stratum Detection Probability: Worst Case Scenarios," submitted concurrently to *Proc. of the INMM & ESARDA Joint Annual Meeting*, 2021.
8. Bevill, A. M., *et al.*, "Multi-stratum Detection Probability Calculations for IAEA Safeguards: Foundations and Early Progress," submitted concurrently to *Proc. of the INMM & ESARDA Joint Annual Meeting*, 2021