DATA ANALYSIS IN SUPPORT OF RADIATION PORTAL MONITORING

Tom Burr [1,*] & Michael S. Hamada[2]
[*]Corresponding Author :Statistical Sciences Group, Los Alamos
National Laboratory, Los Alamos NM, USA

ABSTRACT

Passive gamma and neutron detectors screen for illicit special nuclear material in vehicles crossing borders between countries. This paper illustrates how statistical analysis of archived detector data can help to evaluate special nuclear material detection probabilities and to investigate several issues, including drifting background, background gamma suppression, nuisance alarms arising from either naturally occurring radioactive material (NORM) or cosmic ray bursts, and detector quality control. Statistical techniques described include data smoothing, cosmic ray filtering of neutron alarms, quantile estimation, and pattern recognition. New data analysis of gamma energy spectra suggests that NORM recognition using deployed detectors is difficult.

Keywords: *nuisance alarms, pattern recognition, radiation portal monitors.*

1. INTRODUCTION

Data from passive radiation portal monitors (RPMs) have been collected at various ports of entry since 2002 [1]. The main purpose is to detect potentially harmful radioactive cargo (special nuclear material, SNM) that emits gamma rays and/or neutrons. In the data we analyze below, each vehicle slowly passes by a set of fixed radiation detectors, resulting in a profile time-series measurement from each detector. The most common detector configuration is both a driver's and passenger's side top and bottom panel, each having a neutron count and a low and high energy gamma count recorded every 0.1 second during the vehicle profile. This results in a total of 12 counts (low-energy gamma, high-energy gamma, and neutron counts from each of 4 panels) every 0.1 second. The coarse binning of gamma counts into low or high energy counts is sometimes replaced with a less coarse binning into approximately eight energy bins, and debates continue regarding the pros and cons of deploying higher resolution gamma detectors in primary screening [2]. The data analyses we present use either two-window or eight-window gamma detectors.

This paper illustrates how statistical analysis of archived data can help evaluate special nuclear material (SNM) detection probabilities (DP) and to investigate several issues, including: (1) drifting background; (2) background gamma suppression; (3) nuisance gamma alarms arising from naturally occurring radiation (NORM) and cosmic rays, and (4) the state of detector health, (i.e., detector quality control). In addition, new data analysis results are presented that raise caution regarding a possible option to distinguish NORM from SNM. Statistical techniques described include data smoothing, cosmic ray filtering of neutron alarms, quantile estimation, and pattern recognition.

The following section describes issues 1-4 and Section 3 describes statistical techniques used. Section 4 presents new analysis of gamma energy spectra which suggests that NORM recognition using deployed detectors is difficult. Section 5 is a summary.

2. BACKGROUND

Detection of illicit SNM using passive detectors is complicated by several factors including drifting background [3-5], background suppression due to vehicle self-shielding [6,7], nuisance alarms due to NORM and cosmic rays [8], and the need to monitor detector health.

2.1 Drifting Background

Drifting background requires frequent re-estimation of the background. For example, the background neutron count rate μ should be re-estimated either before each profile or on some comparable frequency, and it is of interest to evaluate the impact of having to estimate the background mean rate μ. Figure 1 plots the actual false alarm probability versus the background averaging period for neutron counts, assuming the counts are Poisson-distributed and that the alarm rule is a simple Shewhart-rule ("maximum count rule") which alarms if the maximum of the n counts exceeds a threshold T (n ranges from approximately 30 to 300 for 3 to 30 second profiles). Figure 1 indicates that approximately 20 seconds of background data are required for there to be negligible effect of having to estimate μ when $\mu = 0.24$ as in our case.

It would not be practical to require 20 seconds of background data to be collected between each vehicle, so instead a moving average is used. The moving average estimate for vehicle *i* is $\hat{\mu}_i = \frac{\sum_{j=i-n+1}^{i} \overline{x}_j}{n}$, where $\overline{x}_j$ is the average count over the leading one-second of profile data that is taken prior to vehicle *j* triggering the vehicle sensor, For example,

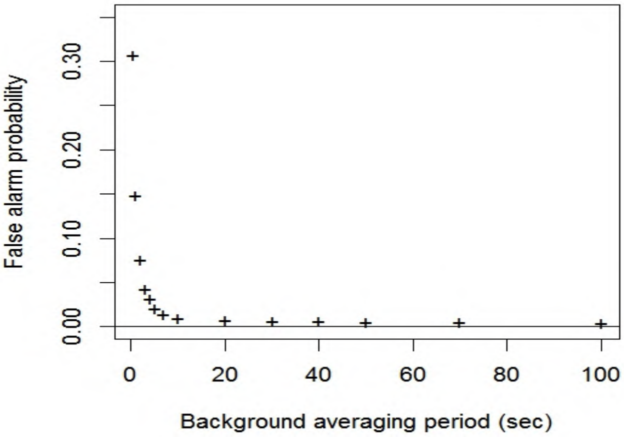


Figure 1. The false alarm probability for the "maximum count" test estimated using 10^6 simulations for background averaging periods ranging from 5 to 1000 (0.5 seconds to 100 seconds). The horizontal line at 0.001 is the theoretical false alarm probability when the background mean is known exactly.

Many fielded detectors use the most recent *n* = 20 pre-profile backgrounds. Because the true neutron background rate μ drifts slowing over time, it is typical to avoid using very large values *n* that would correspond to clock times of one hour or more.

To generate Figure 1, we assumed *n* = 150 for all profiles and a background rate of μ = 2.4 cps that is known, which as mentioned above is a typical neutron count rate [3,4,5]. The value of the threshold *T* that corresponds to a 0.001 false alarm probability, is then the quantile *q* of the Poisson(μ = 2.4) distribution corresponding to a probability of 0.999^{150} . We then used simulation in *R* [9] as follows. For each simulated profile of length *n* = 150, we estimated the background rate μ using a range of background averaging periods and reported the observed FAP at each averaging period, by evaluating the probability that the maximum of the statistic $S = \frac{x - \hat{\mu}}{\hat{\mu}}$ computed over *n* = 150 independent Poisson(μ) random variables *x* exceeds *T.*

2.2 Background Suppression

Vehicle self-shielding implies that vehicles with or without radioactive material will suppress the natural background that typically arises mostly from the asphalt, concrete, air, and rock near the RPM. Although neutron counts exhibit very small amounts of suppression [3,4,8], gamma counts exhibit significant suppression effects [10-12], thus raising the question of how best to mitigate suppression effects for gamma counts. There is a lot of variation in background gamma suppression shape and magnitude from vehicle-to-vehicle. Some reasons for such variation include variation in vehicle sizes, density, and speed.

Several empirical studies [10-12] investigate the suppression-of-gamma-background effect of a typical vehicle. Because vehicle speeds vary, the lengths of vehicle profiles vary from approximately 30 to 300 counts, representing 3 to 30 seconds. For plotting and some analyses, profiles are aligned (stretched or shrunk) to a representative length, such as 150 [12,13]. Several options to adjust for background suppression have been evaluated [10-13]. However, simply subtracting the average background suppression (the "template," with alignment to adjust for unequal profile lengths) results in undesirable patterns in the residuals. One advantage of monitoring count ratios rather than gross counts is that their suppression is less [10,14].

In the top eight plots, which are all gamma counts, Figure 2 illustrates background suppression for an example vehicle profile. The smooth curve (a local kernel smooth fit to the counts using `lokerns` in `R`) shows roughly a

"W-shape" trend that has been predicted using a particle transport model [15] which suggests that suppression occurs due to displacement of the air from which background gammas arise, and from shielding by the vehicle of nearby ground sources such as asphalt. Profile suppression can be defined qualitatively as having an average count that is less than the recent background count. The vehicle profiles we examine throughout are believed to have no SNM and therefore all exhibit suppression.

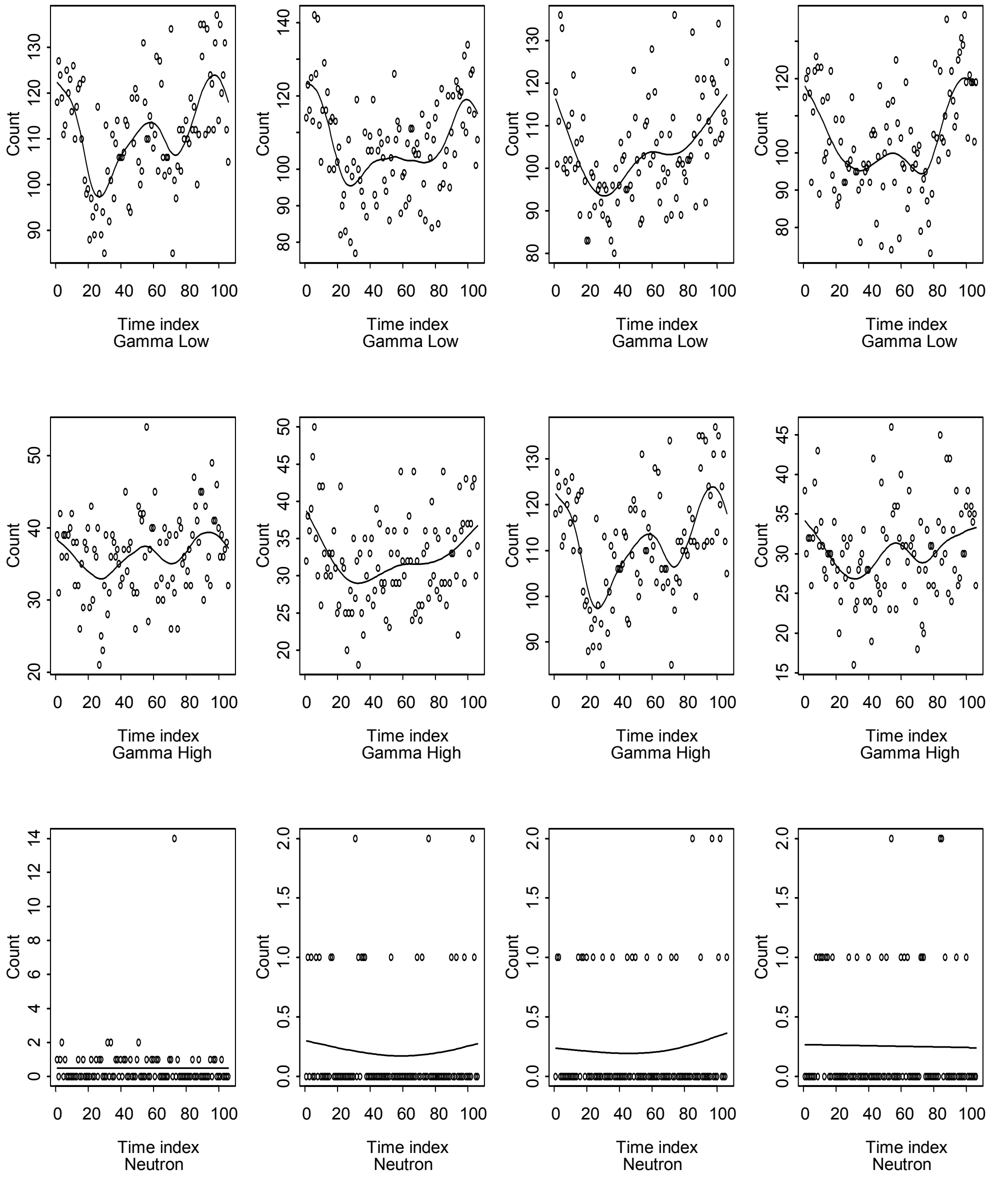


Figure 2. Counts from all 12 panels in an example profile exhibiting a large neutron count (14) in one of four panels. Smooth fits through the data are also shown, and baseline suppression is evident in the gamma detector counts.

2.3 Nuisance Alarms Due To Norm

Nuisance alarms due to NORM cargo or cosmic rays are currently sent to secondary screening, which is more time consuming than primary screening, and includes x-rays, higher-resolution gamma detectors, and isotope identification efforts [16]. The rate of NORM alarms is approximately 1% to 2% of all screened vehicles. The rate of cosmic ray alarms is much lower, on the order of 0.01%, simply because cosmic ray events are relatively rare.

Nuisance alarms due to NORM limit DP for threats. Strategies to recognize common NORM such as cat litter or ceramics depend on the detector energy resolution.

2.4 Nuisance Alarms Due To Cosmic Rays

In contrast to gamma counting, there are very few innocent naturally occurring neutron sources typically found in personal or commercial vehicles. However, cosmic ray burst events (probably due to the so-called ship effect [8]) do produce high neutron counts over time durations that are much shorter than expected from contraband neutron sources in the vehicles. We therefore evaluated a cosmic ray filter (CRF) that filters out any very short duration elevated neutron counts. This filtering has a predictable reduction on the false alarm rate. The CRF [4,17] recognizes that cosmic-ray-induced neutron count bursts have a very short duration, of at most approximately 0.1 seconds. This is in contrast to true neutron sources that have a longer duration, perhaps 1 to 5 seconds, depending on the vehicle profile duration [3,4,8].

Kouzes et al. [8] evaluated the same type of RPM neutron data as in [4], but from another site and only in the context of evaluating whether cosmic ray neutrons arising from the "ship effect" might be recognized and removed. Most of the neutron background arises from a steady rate of cosmic ray induced neutrons. However, it is well known that high density materials (such as the iron and steel in ships or high-density material in cargo vehicles) can enhance the production of cosmic ray neutrons via interactions commonly referred to as the "ship effect." As an example, the neutron count rate near an air-steel interface can be 25 times that at an air-water interface [8].

It is known that the true counts from a fixed neutron source having constant mean are well approximated by a Poisson distribution. In our context, if the mean count rate μ were constant, then the detected neutron counts should be very well approximated by a Poisson distribution having a mean of approximately $\mu = 0.24$ counts per 0.1 sec. However, Kouzes et al. [8] explain that bursts of neutrons from the ship effect and perhaps other sources are superimposed onto an otherwise fairly stable neutron background. For our data and for similar data analyzed in [3] and [8], the fairly stable neutron background exhibits approximately 1% to 2% variation over 24 hours. And, bursts of neutron counts arising from the ship effect due to high density cargo and/or vehicles will not follow a Poisson distribution.

The measured neutron background in our context can be regarded as a mixture of a Poisson background with occasional bursts, or perhaps as a Poisson distribution having nonconstant mean. However, over the duration of a vehicle occupancy, the drifting background is essentially constant and there are often no neutron bursts. It is therefore of interest to assess whether source detection probability results based on simulated background and source data are acceptable for method comparisons. If not, then simulated or real sources must be added to real background to adequately assess source detection probabilities of various methods. Exploratory data analysis (EDA) is therefore mainly intended here to check for the extent of departures from Poisson behavior.

We note here that a Poisson distribution in the true counts emitted per unit time combined with a Binomial model of detected counts results in a Poisson distribution of detected counts. That is, given the true counts C in a time interval, the simplest effective model for the detected counts x is $x|C \sim \text{Binomial}(C, \varepsilon)$ where ε is the detection efficiency, and $C \sim \text{Poisson}(\mu)$. Here, $x|C$ denotes the conditional distribution of x given C. It is then not obvious, but also not difficult to show that $\text{x}|\mu \sim \text{Poisson}(\mu\varepsilon)$. To simplify notation, we write x for the detected counts and include the efficiency term ε in the term μ, so our basic model is $x \sim \text{Poisson}(\mu)$, where $\mu = 0.24$ for all detectors for all lanes.

We report results here for five EDA tasks in comparing neutron counts from each of 4 detectors for 1800 profiles (from multiple lanes) to corresponding Poisson-simulated counts having the same mean of 0.24 counts per 0.1 seconds. In the real data analyzed, the average length in the 1800 profiles is 126, and each profile records counts from 4 neutron detectors, so the total number of neutron counts is 909,200. Profiles longer than length 200 were truncated to length 200.

2.4.1 Exploratory Data Analysis To Check For Poisson Behavior

1) **Frequency Counts**

In the 1800 profiles, there is a clear (strongly statistically significant) indication of too many 3's, 4's, and 5's and other large counts (one 7, two 8's, one 14, and one 17) that are most likely due to cosmic ray events. More formally, a χ^2 goodness of fit test comparing the observed counts to corresponding Poisson simulated counts strongly indicates non-Poisson behavior arising from the several high counts in the real data. We binned the counts so that all bins had at least 5 counts in order for the goodness of fit test based on the χ^2 distribution to be a good approximation [18]. Also, even if we bin all counts of 3 or higher into a single "≥ 3" bin, then the real data counts are still statistically significantly different from the Poisson simulated data counts.

Table 1 lists the relative frequencies of 0, 1, …, ≥ 4 valued counts in the real and the relative probabilities in corresponding Poisson simulated data.

	0	1	2	3	≥ 4
Real	0.79	0.19	0.02	0.002	0.0003
Simulated	0.79	0.19	0.02	0.002	0.0001

Table 1. Relative frequencies of 0, 1, 2, 3 and ≥ 4 valued counts in the real and the relative probabilities in corresponding Poisson simulated data.

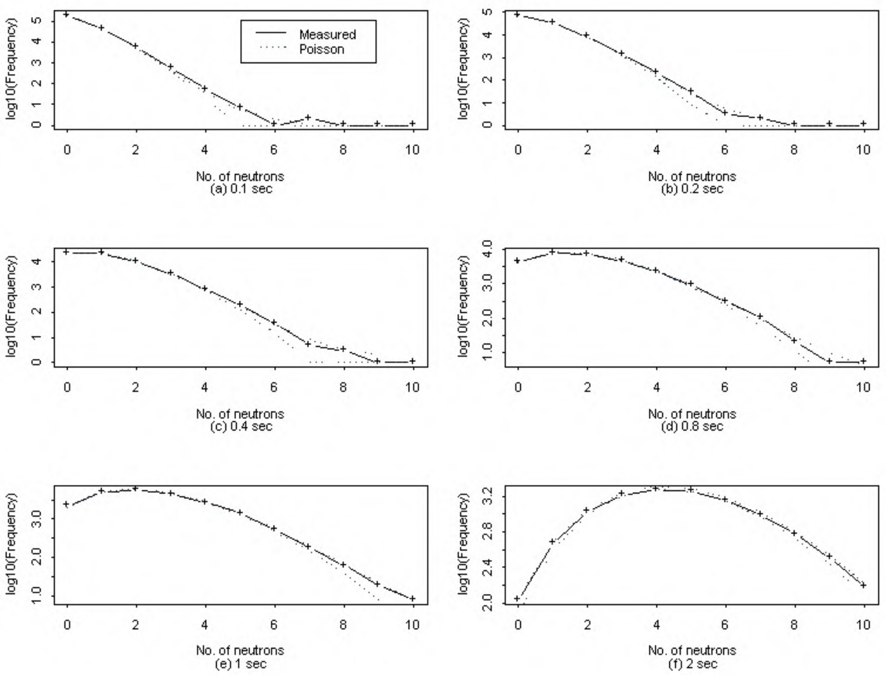
Figure 3 extends Table 1 by also considering nonoverlapping count intervals of 0.2, 0.4, 0.8, 1, and 2 seconds in addition to 0.1 seconds. Figure 3 also includes 95% confidence limits assuming the counts have a Poisson distribution. Because there are 909,200 real neutronscounts in the 1800 profiles (4 neutron detectors per profile, each of length approximately 126) , the 95% limits were estimated using 1000 simulations of 909,200 simulated Poisson counts at 0.1 second, 909,200 counts at 0.2 sec, etc. Notice that the Poisson model is quite good for the longer-duration counts and that it is only mildly violated due to having too many counts of 3 or more. This behavior is consistent with results reported in [8].

Note that elevated rate of counts of 4, 5, 6, and higher in the real data compared to the corresponding counts in Poisson-simulated data could provide a rough estimate of the rate of occurrence of ship-effect neutrons. For example, a rate of 0.0003 (real, observed) in the ≥ 4 bin compared to 0.0001 (simulated) suggests a rate of 1 in 5000 due to ship-effect neutrons. A typical profile is a time series of approximately 125 counts every 0.1 sec for each of 4 neutron detectors, or 500 neutron counts. This implies a ship-effect neutron in approximately every 10 vehicles.

2) Runs Tests

In both the real and simulated data, the average longest-run-above-average length is approximately 2.8. Note that Prob(Pois(0.24) > μ) = Prob(Pois(0.24) > 0) = 0.21, so that a run of length 2 occurs at a given time index within a profile with probability approximately $0.21^2 = 0.04$ and of length 3 with probability approximately $0.21^3 = 0.01$.

For the real and simulated data, the relative frequency of the longest run above the average is given in Table 2. Table 2 also lists an analytical approximation using the Chen-Stein method to the longest run above the average count rate. The Chen-Stein method is described in Section 3.2.



(a) 0.1 sec
(b) 0.2 sec
(c) 0.4 sec
(d) 0.8 sec
(e) 1 sec
(f) 2 sec

Figure 3. The observed number of neutron counts at 0.1, 0.2, 0.4, 0.8, 1, and 2 seconds. One thousand sets of simulated Poisson counts having the same mean were used to generate the 95% confidence limits (solid lines) shown.

Table 2. Relative frequencies of the longest run above the average in the real data, corresponding Poisson-simulated data, and via the Chen-Stein approximation.

	0	1	2	3	4	5	≥ 6
Real	0	0.02	0.38	0.42	0.14	0.03	0.006
Simulated	0	0.02	0.39	0.41	0.14	0.03	0.008
Chen Stein	0	0.01	0.37	0.44	0.14	0.03	0.007

The relative frequencies in Table 2 are very similar for the real and simulated data A χ^2 goodness of fit does detect small yet statistically significant differences between the real and simulated data in run length behavior. However, even very small differences could mean that the CRF would behave differently in the real and corresponding simulated data

3) Serial Dependence Tests

The runs test just described is a type of serial dependence test. There is no serial dependence in the simulated data because the simulation assumed independent and identically distributed Poisson(μ = 0.24) counts. Another serial dependence test is the autocorrelation test which computes the serial correlation r_l at lag 1, 2, … $\text{Lag}_{\max}$, defined as

$$r_l = \frac{\sum_{i=1}^{n-l}(x_{i+l}-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

where x_i is the count at index i. This autocorrelation test also did not indicate serial dependence.

4) cross-talk tests

It is possible that some nuisance alarms in a given lane are caused by nuisance alarms in other lanes. We checked for lane cross talk by evaluating conditional mean count rates. The conditional count rate in a given lane was calculated when the count rate was high in any other lane. Because the conditional count rate in a given lane was statistically indistinguishable from the unconditional count rate, there was no indication of lane cross talk. Specifically, the conditional mean count rate in any lane was statistically indistinguishable from 0.24 cps regardless of the count rate in other lanes. Also, the large neutron counts such as 14 and 17 in a given lane were not temporally associated with large counts in other lanes. And, the location of the maximum count in a given lane was not associated with the location of the maximum count in any other lane.

EDA tasks 1 and 2 both suggest some departure from Poisson behavior in the detected neutron counts. However, there are 909,200 neutron counts for the 1800 profiles each recording neutron counts from 4 detector panels over approximately 12.6 seconds on average. This is a sufficiently large sample size that even small departures from Poisson behavior will be detected as strongly significant. Therefore, as a “bottom-line” comparison of real data and simulated Poisson data, it is also desirable to evaluate the statistical tests for signals on both real and corresponding simulated data.

5) Comparison with corresponding poisson-simulated data

For the 1800 profiles, we simulated a corresponding 1800 profiles and compared the false alarm probability (FAP) for given thresholds and corresponding DP for the two types of injected signals. If real profile *i* had length 120 for example, then corresponding simulated profile *i* also had length 120. We found that the FAP’s and corresponding DP’s are very similar for the real and corresponding simulated data. For example, in 100 sets of simulations of 1800 profiles used to compare to the 1800 real profiles, for the same alarm thresholds, the false alarm rates from the maximum count rule were 0.01 for the real data and 0.01 for the corresponding simulated data. Excellent agreement was observed whether the CRF was used or not. Standard *t*-tests comparing alarm rates consistently showed no statistical difference between real and corresponding simulated data. In addition, we randomly permuted the order in the real neutron time series and found no effect.

These five EDA tasks imply that detection probability results from a study such as [4] that adds simulated source effects onto Poisson-simulated background will closely mimic those from a study that uses real background counts.

Recall that because the background neutron count rate should be re-estimated either before each profile or on some comparable frequency, it is of interest to evaluate the impact of having to estimate the background mean rate μ as was shown in Figure 1. Figure 4 plots neutron counts at various levels of data aggregation for the same real example profile as in Figure 2, showing that short-duration cosmic ray bursts can be filtered out. Recall that Figure 2 plots the low-energy and high-energy gamma counts and the neutron counts for all four detectors for one real profile. Only one of the four neutron panels exhibits a large count of 14 during the profile. Although beyond our scope, another type of statistical filter could require that multiple detectors alarm during the profile, so this example profile would not alarm, despite the single large neutron count from one panel. Notice that the gamma counts in Figure 2 all exhibit baseline suppression, which is more evident in the smooth fits through the count data.

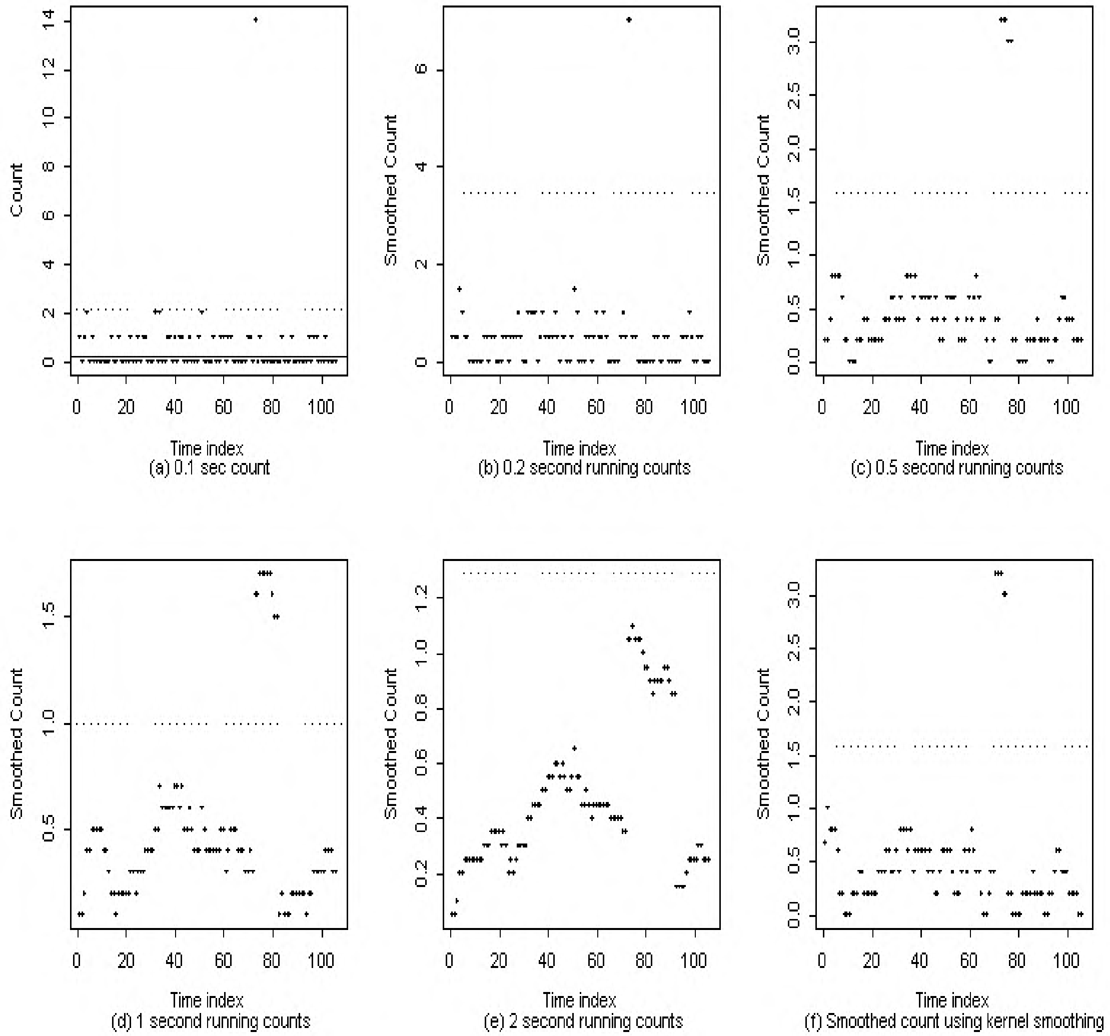


Figure 4. Example real neutron alarm in a single detector (a) original data in the form of counts every 0.1 second; (b)-(e) same as (a), but for average counts over successive overlapping periods of 0.2, 0.5, 1, or 2 seconds; (f) smoothed neutron counts using kernel smoothing. In all except (e), the data alarms at a threshold of median(*x*) + 4 *x* mad(x).

2.4 Detector Health

Although detector health can be monitored using periodic check-source measurements, because the unshielded background changes over time due to environmental changes, archived data is a potential quality control (QC) component to flag measurement anomalies. One QC option is to monitor count ratios from different detector panels such as the low energy gamma counts from panels 1 and 2. The main assumption we then make is that as the background drifts, both detectors will drift in the same direction while if a detector drifts, there will be a growing discrepancy among detectors. If multiple detectors drift simultaneously in the same direction, this is a fundamental ambiguity that cannot in general be distinguished from background drift.

As an example, using training data from December 1-15, 2003 from one site consisting of the low energy gamma counts from panels 1 and 2, and testing data from January-March of 2005, a nominal 1% false alarm rate derived from selecting ratio alarm thresholds from the training data had an actual false alarm rate of 1% to 40% in the testing data, thus indicating some type of detector drifting.

3. STATISTICAL TECHNIQUES

Section two described four issues that impact RPM performance. This section describes in more detail some of the statistical techniques that were used to address the four issues. Statistical techniques described include: data smoothing, cosmic ray filtering of neutron alarms, quantile estimation, and pattern recognition.

3.1 Data Smoothing

There are several good options to smooth time series of counts or spectral data from higher resolution detectors in secondary screening. One common objection to smoothing is that smoothing introduces a bias in the peaks and valleys. Burr et al. [19] review several smoothing options and introduce a multiplicative bias adjustment to mitigate bias introduced by smoothers in peaks and valleys. Burr et al. [20] illustrate that smoothing can increase detection probabilities for injected threats in real background gamma data. Burr and Hamada [4] illustrate that smoothing can increase detection probabilities for injected threats in real background neutron data for some alarm rules. And, qualitatively, Figure 2 illustrated gamma background suppression can be made more obvious by using a simple data smoother such as a moving average.

3.2 The Chen-Stein Approximation To Evaluate Cosmic Ray Filtering

Recall that cosmic ray burst events produce high neutron counts over time durations that are much shorter than expected from contraband neutron sources in the vehicles. We therefore evaluated a cosmic ray filter (CRF) that filters out any very short duration elevated neutron counts. This filtering has a predictable reduction on the false alarm rate. The CRF [4,17] recognizes that ship-effect-induced neutron count bursts have a very short duration, of at most approximately 0.1 seconds. This is in contrast to true neutron sources that will have a longer duration, perhaps 1 to 5 seconds depending on the vehicle profile duration [3,4,8].

The CRF can be defined as follows. For the CRF with the maximum count rule (MCR), if the maximum neutron count during a profile exceeds $T\hat{\sigma}_{\text{bkg}}$, where *T* is a threshold and $\hat{\sigma}_{\text{bkg}}$ is the estimated background count rate, and if a "run" of 3 "large" counts occurs either anywhere in profile or within a time window of the maximum count-alarm index, then the MCR + CRF rule alarms. The threshold *T* is selected via simulation to achieve a desired low false alarm probability (FAP). The "run" length requirement can be modified to any reasonable length such as 2, 3, or 4. In general, the definition of "large" can also be modified, but the low-count rate situations considered here all correspond to a "neutron count of 1 or more" being a "large" count. For any alarm rules having an alarm index (such as the MCR), there is a window of opportunity for the CRF to alarm near the method alarm index. For alarm rules that do not have an alarm index (such as the test that simply uses the average count rate over the entire profile), the CRF must alarm anywhere in the profile.

Currently there are several proprietary versions of CRF's, for example from vendors such as SAIC, Ludlum, and TSA. We will describe two "typical" CRF's. In version one, the CRF must alarm within a time window of the primary alarm rule alarm time. In version two, the CRF can alarm anytime within the profile.

Version two is amenable to a clever approximation (the Chen-Stein method [21]) described as follows. The task is to estimate the probability of a run of various lengths of "above average" counts. In our context with 0.1 second intervals, with $\mu = 0.24$, a count of 1 or more is above average. Denote the longest run of "above average counts" in a series of length *n* as R_n. The well-known Erdos-Renya result is $\lim_{n\to\infty} \frac{R_n}{\log_{(1/p)} n} = 1$ where *p* is the probability of "success," with "success" defined here as an above average count. Because this is an asymptotic result, it is not particularly helpful in calculating the probability that R_n exceeds a specific value such as 3, 4, or 5 for a particular value of *n*.

A result that is less well known involves a remarkably accurate Poisson approximation based on a "Poisson clumping heuristic [22]." The clumping heuristic shows that the number of clumps (a clump is a run of above average counts in our context) will be approximately Poisson in distribution because clump locations have a Poisson distribution. Clump size has a geometric distribution for size *k* occurring with probability $p^{k-1}(1-p)$ and the Chen-Stein theorem can be applied to give the desired approximation with relatively small error bounds. The details will be omitted here, but the Chen-Stein theorem implies the probability that $R_n \geq t$ satisfies

$\left|P(R_n \geq t) - (1 - e^{-\lambda})\right| \leq b_1 \min(1, \frac{1}{\lambda})$, where $\lambda = p^t + (n-t)(1-p)p^t$.

The term b_1 satisfies $b_1 \le (2t+1)((1-p)p^t)^2\{n-t+(1-p)^{-1}+p\{(1-p)^2(2t+1)\}^{-1}\}$. The term *p* in the previous expression is given by $p = \text{Prob}(\text{Pois}(\mu) > \mu) = 1 - e^{-\mu}$ for μ in (0,1) as we have here with μ = 0.24. This approximation was used in row 3 of Table 2.

One additional exploratory data analysis is whether there is a trend in the CRF alarm rate as a function of averaging period. Figure 5 indicates that there is no noticeable trend in the CRF as a function of averaging period.

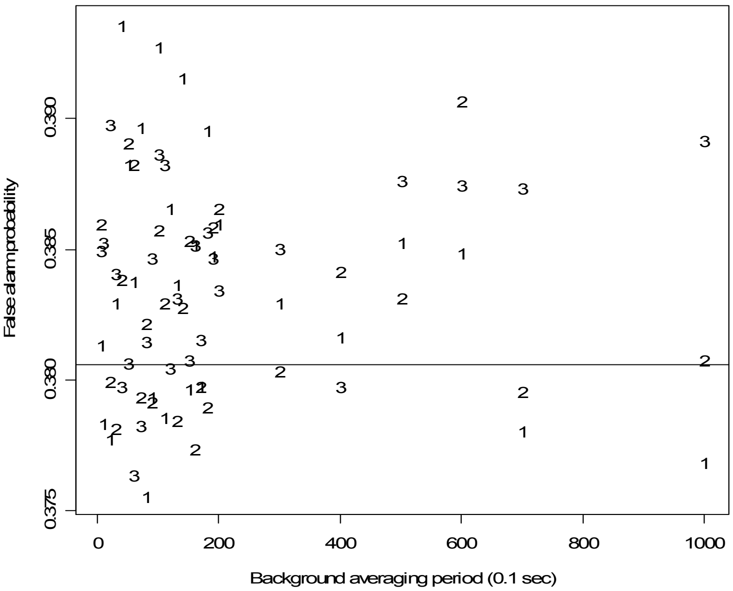


Figure 5. Checking for a trend in the CRF alarm rate as function of averaging period. Points marked with a "1" are from set 1 of 10,000 simulations, thosed marked "2" are from set 2, and those marked "3" are from set 3. The horizontal line at 0.382 is the theoretical approximate CRF alarm rate based on the Poisson clumping approximation. There is no noticeable trend.

As an example calculation, for a profile of length 100, probability p_3 of observing at least one run of 3 values above the mean could be estimated by assuming there are 98 possible starting indices for a run of three or more nonzero counts. Therefore, the probability could be estimated as 1-(1-P(Poisson(x>0|μ=0.24))^3)^{98} = 0.62, assuming there are 98 independent tries for a run of length 3. However, the event of a run of length 3 starting at index *i* is not independent of the event of a run of length 3 starting at index *i*+1. Therefore, this is a rough approximation that will overestimate the true probability. At the other extreme, one could consider only the 33 nonoverlapping sets of 3 indices and use the estimate 1-(1-P(Poisson(x>0|μ=0.24))^3)^{98} = 0.18. The Chen-Stein approximation adjusts for the dependence just described and the estimate is 0.53. Table 3 compares the Chen-Stein and naïve approximations to the true probability (estimated using 10^4 simulations).

Table 3. Probability of a run of at least length 2, 3, or 4. Notice the excellent agreement with the Chen Stein approximation. The naïve low estimate is based on assuming 99, 98, or 97 independent tries for a run of length 2, 3, or 4. The naïve high estimate is based on assuming 50, 33, or 25 independent tries for a run of length 2, 3, or 4.

	2	3	4
Chen Stein	0.97	0.53	0.15
Naïve low	0.81	0.18	0.03
Naïve high	0.99	0.62	0.18
True	0.97	0.53	0.15

The Chen-Stein approximation can be used together with alarm probabilities for the MCR for example, to approximate the FAP of the combined CRF and MCR rule. However, the CRF and MCR rules are not independent, only nearly so. We have found for example that the FAP of the combined rule is very well approximated by the product of the FAPs of the CRF and the MCR.

3.3 Quantile Estimation

Suppose the FAP of many candidate alarm rules under several conditions (background count rates, vehicle speed, vehicle background suppression magnitude, etc.) is to be evaluated in off-line analyses, and that the FAP should be small, such as 0.001. For any alarm rule that requires simulation (such as sequential testing methods) to estimate the FAP, the "brute force" approach evaluates trial threshold values many times. For example, reliable estimation of the 0.001 quantile requires tens of thousands of simulations. To estimate the FAP, written formally as $\text{FAP} = \int_x I(S(x,\theta) > q) f(x)dx$ where the indicator function *I*() = 1 if its argument is true, *S*(*x*,θ) is the test statistic which depends on the data *x* and parameters θ, *q* is the alarm threshold (quantile), and *f*(*x*) is the probability distribution of the data *x*.

Alternatively, importance sampling can be useful in reducing computational time in the context of extreme quantile estimation. The basic idea is to express the FAP as $\text{FAP} = \int_x I(S(x,\theta) > q) \frac{f(x)}{g(x)} g(x)dx$ and choose a reasonable importance sampling function *g*(*x*) that never vanishes over the range of *x* (to avoid division by 0 in the term $\frac{f(x)}{g(x)}$), that is fast to simulate from, and that leads to small variance in the estimated quantile *q*. Picard et al. [5] provide detailed examples of importance sampling for quantile estimation using sequential testing in RPMs.

3.4 Pattern Recognition

In a typical pattern recognition problem, the data consist of *n* cases of (*y*, *X*) pairs where the integer $y \in (1, 2,..., J)$ is the class and *X* is a *p*-dimensional predictor vector. The goal is to use *X* to predict the class *y*, and this task is sometimes called classification, discriminant analysis (DA), pattern recognition, or supervised learning. Regarding notation, vectors and scalars can be distinguished by context and definition. The most common pattern recognition data model assumes that a categorical response *y* depends on a fixed-dimension predictor *X*. The pattern recognition task is to estimate *f*(*X*) = Prob(*y* = 1| *X*). The most well studied version of this task assumes the following: (1) all components of *X* are real-valued; (2) *X* has fixed dimension, and (3) training cases consisting of (*X, y*) pairs are independent.

In the RPM context, pattern recognition can be used to recognize common NORM cargo (and perhaps not always send NORM-like cargo to secondary screening) and/or to distinguish NORM from threat SNM. To recognize common NORM, one of the best methods using the systems described here (two-energy gamma and neutron) uses a nonparametric density estimation method for pattern recognition [23].

4. CASE STUDY: DATA MINING INFORMS PATTERN RECOGNITION

To recognize common NORM and distinguish common NORM from threat SNM isotopes, reference [24] used the data in Table 4. The data in Table 4 was provided to Dalal and Han [24] by Pacific Northwest National Laboratory as a summary of data from fielded detectors. Table 4 shows the relative frequencies ("signature" relative frequencies") expected in three energy windows for each of seven material categories, including two threat SNM materials (highly enriched uranium and weapons grade plutonium) and five common NORMs.

Reference [24] made a key simplification by assuming that the signature relative frequencies in the three windows did not vary across profiles of the same type of material category. Therefore, variation from profile to profile in the average observed relative frequency of the counts in each of the three bins was due entirely to variation arising from multinomial sampling using profiles having varying lengths, each approximately 150 counts of 0.1 second duration.

Using exploratory data analysis of real profiles carrying cat litter or bentofix, here we modify the simulation results presented in [24]. Our main finding is evidence of non-negligible between-profile variation in the signature relative frequencies in Table 4. For example, Figure 6 illustrates the between-profile variation in the relative frequencies in energy bin 1 across 50 bentofix profiles and 98 cat litter profiles. This between-profile variation is considerably larger than can be explained by multinomial sampling, as explained below.

Figure 7 illustrates (as did [24]) that the ratio of counts is much more consistent across a profile than the low (or gross) energy count, although as explained below the ratio of counts does show evidence of modest within-profile variation beyond that which can be explained by multinomial sampling variation. The ratio is defined as $R = C_L/C_T$,

where C_L is the counts in the low energy bin and C_T is the counts in all energy bins.
Figure 8 plots simulated realizations from the signature relative frequency vectors in Table 4 without (a) and with (b) between-profile variation in the signature.

Table 4. The mean relative frequency in energy windows 1, 2, and 3 for 2 threat SNM materials (HEU and WGPu) and for 5 common NORMs assumed in reference [24].

Material	Window 1	Window 2	Window 3
HEU	0.954	0.033	0.013
Fertilizer	0.635	0.243	0.122
Tile	0.658	0.242	0.100
WGPu	0.934	0.061	0.005
Cat Litter	0.631	0.292	0.077
Road Salt	0.662	0.273	0.065
Background	0.651	0.249	0.100

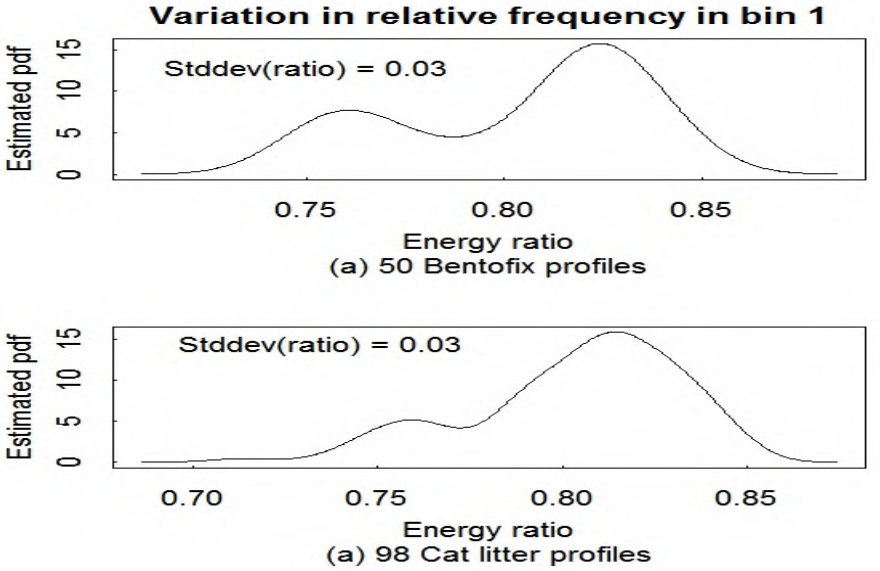


Figure 6. Summary of 50 Bentofix profiles and 98 cat litter profiles, illustrating non-negligible between-profile variation in the ratio of the low energy bin count to the total count.

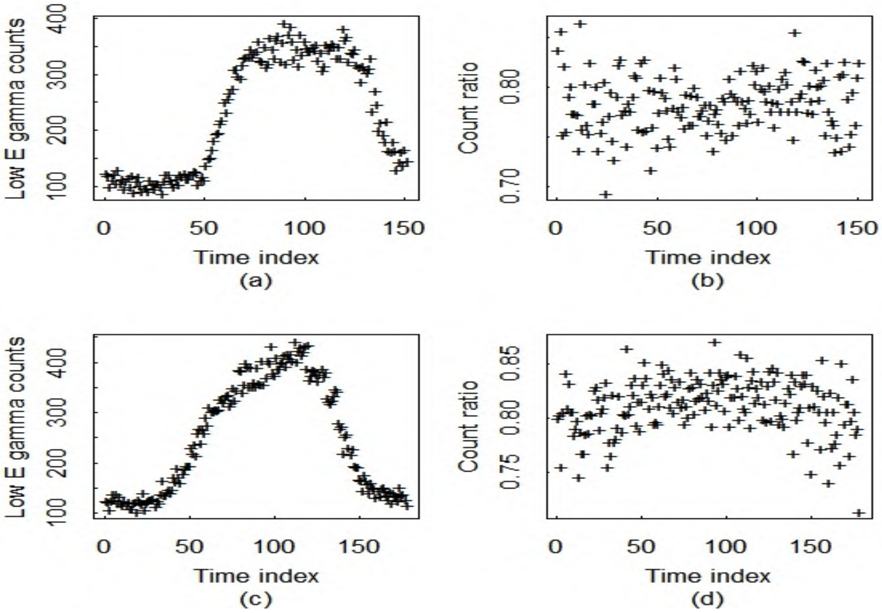


Figure 7. Examples of two NORM-carrying profiles with large low-energy gamma counts, and the corresponding count ratio for both profiles.

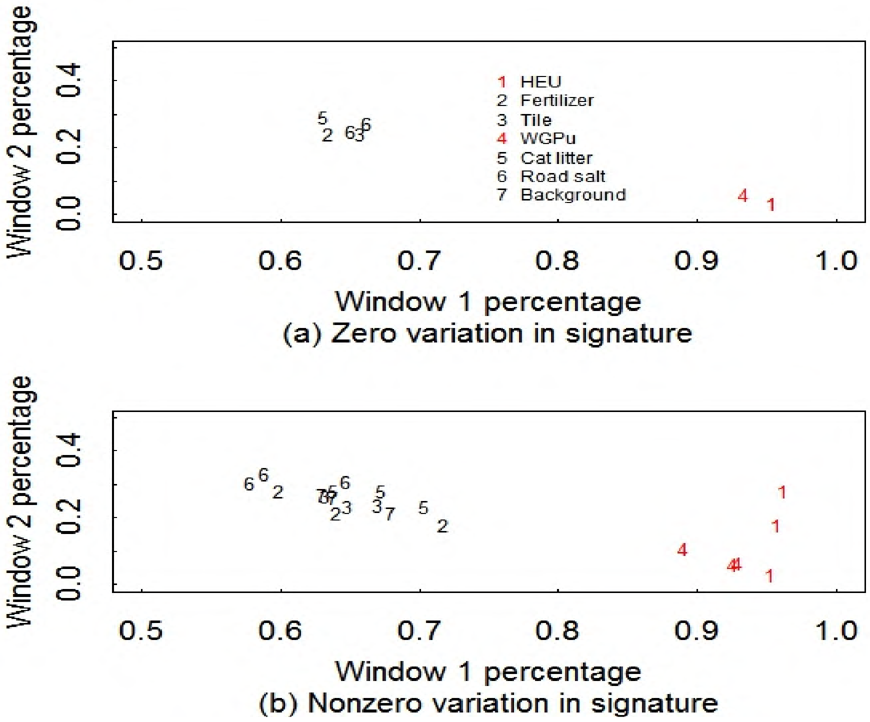


Figure 8. Window 2 observed percentage versus window 1 observed percentage for the 7 material categories in Table 4.

Guided by exploratory data analysis, we modify the results in [24] as we explain next. The three key assumptions by Dalal and Han [24] are:

Assumption 1) The count rate $C_{b,d} \,|(\text{class} = k, d, M) \sim \text{Poisson}(g(d,M)\lambda_{b,k})$ in energy bin *b* at distance *d* from the detector for *M* grams of material in class *k*, where $g(d,M)$ is an attenuation factor and $\lambda_{b,k}$ is the gamma emission rate for a unit quantity of material *k* in energy window *b* at distance *d* = 0. Let $N_d = \sum_b C_{b,d}$ denote the sum of counts over all bins at distance *d* (the time step in the profile is a surrogate for the distance *d*). Assuming that the gross counts N_d have a Poisson distribution, then conditional on N_d, the counts $C_{b,d}$ for a given *k* have a multinomial distribution, denoted $C_{b,d} \sim \text{Multinomial}(N_d; p_{1,k}, p_{2,k}, p_{3,k})$ where $p_{b,k} = \frac{\lambda_{b,k}}{\sum_b \lambda_{b,k}}$. Dalal and Han [24] then used maximum likelihood applied to the multinomial distribution to implement what was called a Bayesian classifier. It was called "Bayesian" because it relied on knowing the prior probabilities of each material category. Under the assumption of equal prior probabilities of each category, it is a maximum likelihood classifier based on the multinomial distribution.

Assumption 2) The attenuation factor *g*(*d*,*M*) does not depend on energy bin. This assumption is not true in general because gamma counts at different energy bins are attenuated more or less as function of shielding inhomogeneities. Therefore, this assumption might limit the applicability of the low misclassification rates reported by [24] to vehicles having homogeneous shielding effects, such as might be found in experimental profiles. And, this assumption is the leading candidate explanation for why our real data exhibits profile-to-profile variation beyond that which is explainable from multinomial sampling alone.

Assumption 3) A Dirichlet prior was used to describe how well the $\lambda_{b,k}$ values were estimated as a function of the number of evaluated profiles of each material type.

The next three subsection describe our modifications to the estimated misclassification rates arising from assumptions 1-3.

4.1 Exploratory Data Analysis Reveals Large Profile-To-Profile Variation In Relative Count Rates

Exploratory data analysis of 98 cat litter profiles and 50 bentofix (a clay-based commercial lining product) profiles indicates considerable within-profile and between-profile variation in the relative count rates in the different energy bins. For example, the data for Figure 7 shows a standard deviation across the 98 or 50 bentofix profiles of the average (across the time series in each profile) ratio R_{low} of the low energy bin counts to the total counts of approximately 0.03. The 0.03 between-profile standard deviation in the average ratio is much too large to occur by chance, as we confirmed by simulation in R [9] by generating a reference distribution for comparing to 0.03. The reference distribution was generated for 98 or 50 synthetic profiles, with each set generated 10^6 times using multinomial sampling as the only source of variation from profile to profile. In the 10^6 simulated sets of 98 cat litter profiles, the average standard deviation in the mean ratio across the 10^6 simulated data sets of 98 profiles was 0.002, with a 99.9% upper quantile of 0.021 (and corresponding results are approximately the same for the 50 bentofix profiles). Therefore the 0.03 between-profile standard deviation in the average ratio is much too large to occur by chance. Similarly, in the 10^6 simulated sets of 98 cat litter profiles, the average standard deviation in the within-profile count ratio across the 10^6 simulated data sets of 98 profiles was 0.002, with a 99.9% upper quantile of 0.024 (and corresponding results are approximately the same for the 50 bentofix profiles). The observed within-profile variation in R_{low} across time points is approximately 0.024 so the simulated reference distribution in R illustrates that the within-profile variation is at the 99.9% quantile. Therefore, 0.024 is not as extreme a tail event as the between-profile standard deviation in the mean count rate, but it corresponds to a 0.001 tail area. Therefore, there is statistical evidence of a modest over-dispersion of the within-profile ratio, and there is strong serial correlation of the ratio (with a lag-1 autocorrelation of approximately 0.97 for both cat litter and bentofix) within each profile.

Because our focus is on the average ratio over the profile, we are more concerned with large (0.03) between-profile variation in the average ratio than with the within-profile variation of the ratio, but the within-profile variation analysis is included for completeness.

4.2 Application Of The Dirichlet Distribution

To address the between-profile variation in R_{low},we invoke the Dirichlet distribution, but in a different role than that used in [24]. Our interpretation of the role for the Dirichlet distribution is that each vehicle carrying the same cargo type (such as cat litter) has a unique mean count rate. As shown in Section 4.1, this mean count rate varies during the profile, presumably due to spatially inhomogeneous shielding effects that impact the count rates at different energies in different way (which as mentioned above can be modeled as variation in the attenuation factor *g*(*d*,*M*)). That is, the Dirichlet distribution can model cross-vehicle and cross-time variation within a vehicle.

Our model as informed by exploratory data analysis modifies the model in [24] so that the joint distribution of $\{X_1, X_2, ..., X_k\}$ is multinomial$(p_1, p_2, \ldots, p_k)$ where
$(p_1, p_2, \ldots, p_k) \sim$ Dirichlet$(\theta_1, \theta_{2,\ldots,} \theta_k)$, with $(\theta_1, \theta_{2,}\ldots, \theta_k)$ estimated from variances in real data as illustrated in Section 4.1. For simplicity here, we average the counts across time to mitigate the cross-time effects and assume the resulting cross-time effects during a single profile are negligible. Because the true $(p_1, p_2, \ldots, p_k)$ values vary across vehicles that only nominally carry the "same" cargo and shielding effects (when actually there is non-negligible vehicle-to-vehicle variation), we use the $(p_1, p_2, \ldots, p_k) \sim$ Dirichlet$(\theta_1, \theta_{2,}\ldots, \theta_k)$ assumption to capture cross-profile variation within the same material class. In the Dirichlet$(\theta_1, \theta_{2,}\ldots, \theta_k)$ distribution the sum $\theta = \sum_{i=1}^{k} \theta_i$ determines the variance in the vector $(p_1, p_2, \ldots, p_k)$. The sum $\theta = \sum_{i=1}^{k} \theta_i = 30$ corresponds to the 0.03 variance observed in the cat litter and bentofix profiles.
The data model by [24] relies on the well-known fact that if $\{X_1, X_2, ..., X_k\}$ are k independent (but not necessarily identically distributed) Poisson variables, then the joint distribution of $\{X_1, X_2, ..., X_k\}$ conditionally to their sum is a multinomial distribution. But, again, this assumes $g(d,M)$ does not depend on energy bin, which implies that the multinomial distribution has identical parameters across time points and between profiles that carry material in the same material category.

4.3 Modified Misclassification Rates

Using only cross-vehicle variation (assuming a 0.03 standard deviation of the between-profile average ratio of the low energy to total energy count ratio as observed in the bentofix and cat litter profiles), and ignoring within-vehicle variation, we implement the same Bayesian classifier as [24]. Table 5 gives the resulting modified misclassification rate (MCR) estimates. For example, the MCR for cat litter is approximately 50%. These MCR estimates are only slightly higher if we also include cross-time variation of 0.024 as found in the real bentofix and cat litter profiles. The cross-time variation is mitigated because we use the mean ratio across the profile. And a profile typically lasts for 10 to 15 seconds, resulting in 10 to 15 one-second ratios that can be averaged. Note that the standard deviation of the ratio across profiles (approximately 0.03 for the bentofix and cat litter examples) already includes the effects of within-profile variation, so re-doing Table 5 with cross-time and within-profile variation overstates slightly the two sources of variation (within-profile and between-profile). Therefore, we show results just for the case of using only cross-vehicle variation in Table 5, and the results with added cross-time variation are mentioned above for completeness.

Table 5. Confusion matrix with $\theta = \sum_{i=1}^{k} \theta_i = 30$ in the Dirichlet$(\theta_1, \theta_{2,\ldots,} \theta_k)$ distribution with the sum. The true category is in the rows and the inferred category is in the columns.

True/Guess	**1= HEU**	**2 = Fertilizer**	**3 = Tile**	**4 = WGPu**	**5 = Cat litter**	**6 = Road salt**	**7 = Bkg**
1	1266	0	0	170	0	0	0
2	0	868	262	0	95	12	197
3	0	378	533	0	159	152	237
4	188	0	0	1277	0	0	0
5 55% MCR	0	52	59	0	772	375	140
6	0	18	104	0	338	861	74
7	0	394	406	0	187	179	247

To summarize, in the real data analyzed, $\theta \approx 30$ in the Dirichlet distribution, resulting in an estimated misclassification rate of the multinomial-based Bayesian classifier for cat litter of approximately 0.50 as shown in Table 5. By comparison, [24] report zero misclassifications for cat litter and almost no misclassifications for any of the categories except for approximately a 10% mistake rate with tile confused as background or vice versa. Therefore, between-profile variation is an important noise source.

In addition, using the same type of exploratory data analyses, but for other fielded detectors, for 8-window plastic, again $\theta \approx 30$ and for 8-window sodium iodide (NaI has more than 8 windows, but for comparison we binned the sodium iodide detectors to 2-window or 8-window), $\theta \approx 50$ on basis of between and within profile standard deviations. Therefore, the misclassification rate is predicted to be slightly lower for 8-window NaI than for 8-window plastic scintillator detectors, but still considerably higher than reported in [24]. Figure 9 plots the overall misclassification rate obtained by simulation for a range of θ values.

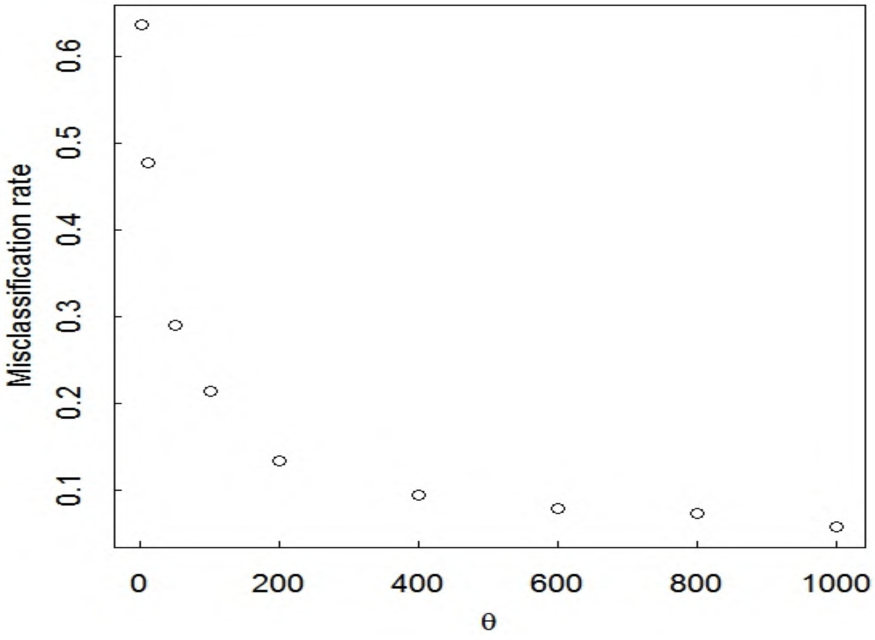


Figure 9. Misclassification rate versus $\theta = \sum_{i=1}^{k} \theta_i$ in the Dirichlet(θ_1, $\theta_{2,\ldots,}$ θ_k) distribution.

5. SUMMARY

This paper illustrated how statistical analysis of archived data can help evaluate special nuclear material (SNM) detection probabilities (DP) and investigated several issues, including: (1) drifting background, (2) background gamma suppression; (3) nuisance gamma alarms arising from naturally occurring radiation (NORM) and cosmic rays, and (4) the state of detector health. In addition, new data analysis results were presented that raise caution regarding a possible option to distinguish NORM from SNM. Statistical techniques described include data smoothing, cosmic ray filtering of neutron alarms, quantile estimation, and pattern recognition.

6. ACKNOWLEDGEMENTS

We acknowledge the Department of Homeland Security for funding this work under DOE Contract Number DE-AC52-06NA25396 for the management and operation of Los Alamos National Laboratory. We thank Mark Abhold and Christopher Lovejoy for insightful comments.

7. REFERENCES

[1]. B. Geelhood, J. Ely, R. Hansen, R. Kouzes, J. Schweppe, R. Warner, Overview of portal monitoring at border crossings, *IEEE Nuclear Science Symposium –Conference Record*, 513-517 (2004).

[2]. National Research Council, Evaluating testing, costs, and benefits of advanced spectroscopic portals for screening cargo at ports of entry: interim report, National Research Council, National Academies Press, ISBN: 0-309-14022-6, http://hps.org/govtrelations/documents/nas_testing-portal-monitors_062509.pdf (2009).

[3]. R. Kouzes, E. Siciliano, J. Ely, P. Keller, R. McConn, Passive neutron detection for interdiction of nuclear material at borders, *Nuclear Instruments and Methods in Physics Research A*, 584, 383-400 2008.

[4]. T. Burr, M.S. Hamada, Moving neutron source detection in radiation portal monitoring, to appear, *Technometrics* (2013).

[5]. R. Picard, T. Burr, M.S. Hamada, Quantile estimation for radiation portal monitoring, to appear *Technometrics* (2013).

[6]. C. Lo Presti, D. Weier, R. Kouzes, J. Schweppe, Baseline suppression of vehicle portal monitor gamma counts: a characterization study, *Nuclear Instruments and Methods in Physics Research A*, 562, 281-297 (2006).

[7]. T. Burr, K. Myers, Background suppression effects on signal estimation, *Applied Radiation and Isotopes*, 67, 1729-1737 (2009).

[8]. R. Kouzes, J. Ely, A. Seifert, E. Siciliano, D. Weier, L. Windsor, M. Woodring, J. Borgardt, E. Buckley, E. Flumerfelt, A. Olilveri, M. Salvitti, Cosmic-ray-induced ship-effect neutron measurements and implications for cargo scanning at borders, *Nuclear Instruments and Methods in Physics Research A*, 587, 89-100 (2008).

[9]. R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.R-project.org (2004).

[10]. T. Burr, J. Gattiker, K. Myers, G. Tompkins, Alarm criteria in radiation portal monitoring, *Applied Radiation and Isotopes*, 65, 569–580 (2007).

[11]. I. Shokair, C. Kunz, Hierarchical cluster analysis of background suppression of RPM data. Sandia National Laboratories report SAND2007-8132 (2007).

[12]. D. Weier, J. Ely, B. O'Brien, B., R. Kouzes, Data-based considerations in portal radiation monitoring of cargo vehicles. 45th Annual meeting of the Inst. of Nuclear Materials Management, CD-ROM, 2004.

[13]. J. Gattiker, T. Burr , Bayesian estimation of the source and suppression effects in vehicle radiation signatures, *Journal of Nucl. Mater. Management,* 37(3), 14-24 (2009).

[14]. J. Ely, R. Kouzes, J. Schweppe, E. Siciliano, D. Strachan, D.Weier, The use of energy windowing to discriminate SNM from NORM in radiation portal monitors, *Nuclear Instruments and Methods in Physics Research A*, 560(2), 373-387 (2005).

[15]. S. Robinson, S. Bender, E. Flumerfelt, C. LoPresti, M. Woodring, Time series evaluation of radiation portal monitor data for point source detection, *IEEE Transactions in Nuclear Science,* 56(6), 3688-3693 (2009).

[16]. T. Burr, M.S. Hamada, Radio-isotope identification algorithms for NaI gamma spectra.*Algorithms* 2(1): 339-360 (2009).

[17]. D. Beddingfield, H. Menlove, Statistical data filtration in neutron coincidence counting. Los Alamos National Laboratory Research Report LA-12451-MS (1992).

[18]. P. Bickel and K. Doksum, Mathematical Statistics Volume 1 (2007).

[19]. T. Burr, N. Hengartner N., E. Matzner-Lober, S. Myers, L. Rouviere, Smoothing low resolution NaI spectra, *IEEE Transactions on Nuclear Science* 57(5), 2831-2840 (2010).

[20]. T. Burr, M.S. Hamada, N. Hengartner, Impact of spectral smoothing on gamma radiation portal alarm probabilities, *Applied Radiation and Isotopes* 69, 1436-1446 (2011).

[21]. R. Arratia, L. Goldstein, L. Gordon, L., Poisson approximation and the Chen-Stein method, *Statistical Science* 5(4), 403-424 (1990).

[22]. M. Waterman, M. Vingron, Sequence comparison significance and Poisson approximation, *Statistical Science* 9 (3), 367–381 (1994).

[23]. T. Burr, J. Doak, Distribution free discriminant analysis, *Intelligent Data Analysis* 11(6): 651-66, (2007).

[24]. S. Dalal, B. Han, Detection of radioactive material entering national ports: a Bayesian approach to radiation portal data, *The Annals of Applied* Statistics 4(3), 1256-1271 (2010).