DATA MINING IN RADIATION PORTAL MONITORING

Tom Burr [1], Michael S. Hamada[1], Kary Myers[1], Nicholas Hengartner[2] & Richard Picard[1]
[1]Statistical Sciences Group, Los Alamos National Laboratory[1], USA
[2]Information Sciences, Los Alamos National Laboratory[2], USA

ABSTRACT

Currently deployed passive gamma and neutron detectors screen for illicit nuclear material. Archived data can help evaluate special nuclear material detection probabilities and investigate several related issues, including (1) nuisance gamma alarms arising from naturally occurring radiation, (2) the impact of drifting neutron and gamma background rates, and (3) radioisotope identification performance. This paper illustrates roles for data mining to investigate issue (1) and briefly reviews data mining to investigate issues (2) and (3).

Keywords: *alarm criteria, passive gamma counting, background suppression, experimental design, sequential testing*

1. INTRODUCTION

Passive radiation portal monitors (RPMs) have been deployed at various screening locations since 2002 [1] to detect potentially harmful radioactive cargo [such as special nuclear material (SNM)] that emits gamma rays and/or neutrons. Current systems use polyvinyl-toluene-based plastic scintillation gamma-ray detectors coupled to photomultiplier tubes. These systems can provide only very coarse energy resolution into a few energy bands, such as low energy and high energy. Alarm rules for the simplest systems are based on the net counts above background for either the low-energy or the high-energy gamma counts or for the neutron counts. Proposed systems use higher-resolution sodium iodide (NaI) detectors for gamma radiation in primary vehicle screening.Because some of the nonthreat cargo contains naturally occurring radioactive material (NORM), such as the potassium in cat litter, the majority of alarms are not due to statistical fluctuations, but instead are true (nuisance) alarms due to NORM [2], [3], [4]. Also, because a simple count criterion leads to many nuisance alarms arising from NORM and because background suppression (see the next section) by the vehicle is smaller for ratios of gamma counts than for counts alone, some systems include both gamma count and gamma count ratio alarm criteria [5], [6]. Following current convention, we define the gamma ratio as the ratio of low-energy gamma counts to total-energy gamma counts, where the total-energy count is the sum of low-energy and high-energy counts. One strategy for nuisance alarms is to define and recognize "signatures" of certain types of NORM so that many alarms can be quickly resolved as being innocent [4]. Burr and Myers [4] considered candidate profile features, such as the peak width and the maximum energy ratio and the use of pattern recognition methods from the machine and statistical learning communities, to illustrate the extent to which several common types of NORM can be distinguished. Currently deployed passive gamma and neutron detectors screen for illicit nuclear material. Archived data can help evaluate SNM detection probabilities (DPs) and investigate several related issues, including (1) nuisance gamma alarms arising from NORM, (2) the impact of drifting neutron and gamma background rates, and (3) radioisotope identification (RIID) performance. This paper illustrates roles for data mining to investigate issue (1) and briefly reviews data mining to investigate issues (2) and (3).The following sections include additional background; features for pattern recognition, pattern recognition methods, and pattern recognition results for issue (1) and then describe roles for other data mining efforts for issues (2) and (3).

2. BACKGROUND

Figure 1 shows one screening location, where four detector panels each record a neutron and a low- and high-energy gamma count every 0.1 seconds for 5–20 seconds, resulting in a 12-component time series of 50 to 200 observations. Nuisance alarms due to NORM limit DPs for threats. Strategies to recognize common NORM, such as cat litter or ceramics, depend on the sensor energy resolution. Figure 2 (reproduced from Burr and Myers [4]) illustrates the extent to which different common NORM categories have a signature using a two-dimensional representation for each profile. The scaled maximum low-energy count rate versus the scaled maximum ratio of the high- to low-energy counts is plotted for each of several profiles of eight NORM categories. One of the best methods using the systems described here (two-energy gamma and neutron in each of four panels) uses a nonparametric density estimation method for pattern recognition [4], [7] applied to tens of features such as those used in Figure 2 derived from the 12-component time series for each profile. Although some common NORMs do appear to have a signature, any vehicle currently having a large count is subject to further investigation in secondary screening, as described in the next paragraph. Such additional investigation results in slower vehicle transit times.

Nuisance alarms due to NORM limit DPs for threats. For gamma detectors, strategies to recognize common NORM, such as cat litter or ceramics, depend on the sensor energy resolution. Burr and Myers [4] also illustrated the extent to which different common NORM categories have a signature with very-low-resolution gamma detectors and use nonparametric density estimation for pattern recognition. Vehicles that alarm in primary screening go to secondary screening, where, for example, higher-resolution gamma and x-ray measurements are made. RIID is a major challenge for low-, medium-, or even high-resolution gamma spectra. On the basis of a small test data set, we have found in unpublished work that medium-resolution detectors, such as the handheld NaI detectors used in secondary screening, appear to be competitive with high-resolution detectors. A current challenge is to evaluate the cost/benefit that medium-resolution NaI detectors might provide in primary screening, deployed as so-called "advanced spectroscopic portals" (ASPs). Testing with NaI detectors is ongoing to estimate low-, medium-, and high-resolution detector performance on several metrics. One straightforward opportunity to improve RIID performance and testing appears to be spectral smoothing, with adjustments made to preserve key spectral regions of interest, such as peak areas [8]. RIID algorithms can be tested for more measurement scenarios by reducing the count time and by augmenting real spectra with realistic synthetic spectra. Model uncertainty will play a key role in assessing the adequacy of synthetic spectra.

Figure 1. Example screening location with four detector panels surrounding the vehicle.

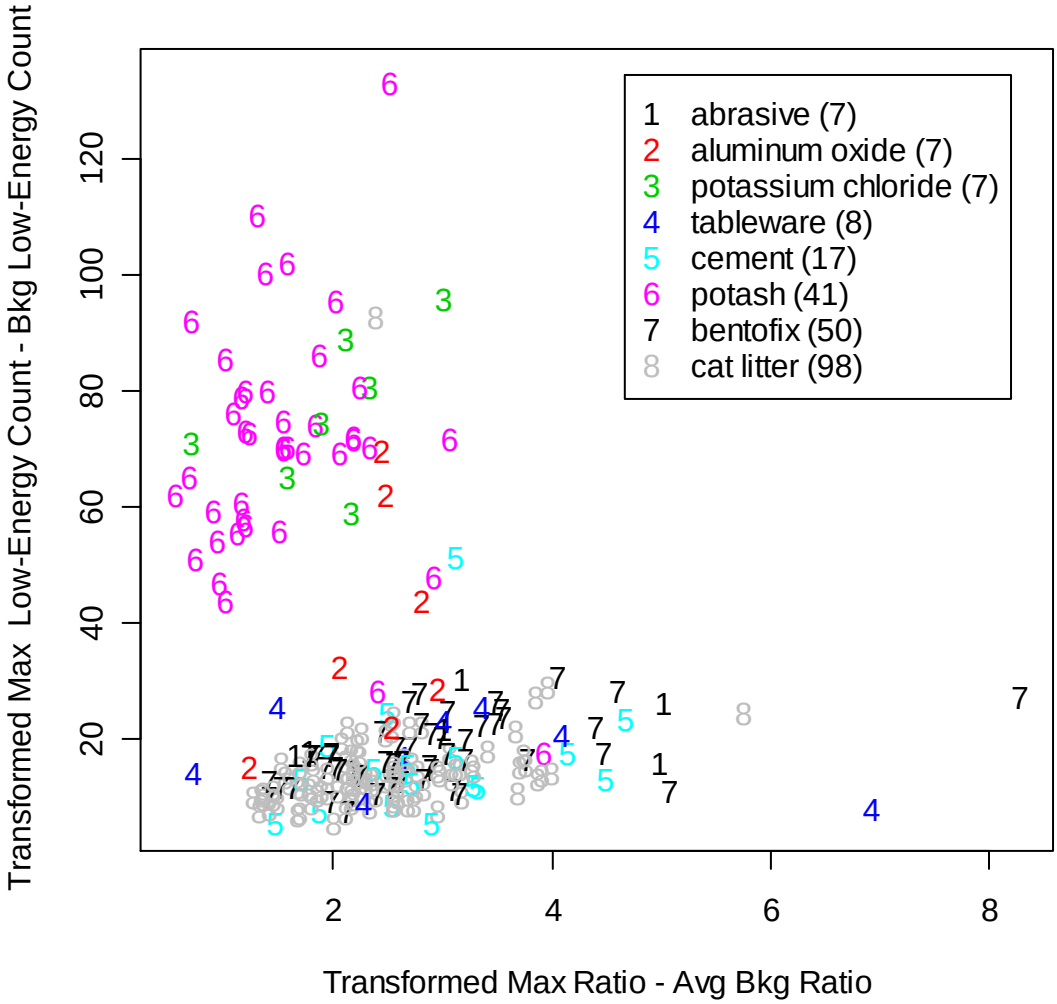

Figure 2. The scaled maximum low-energy gamma count versus the scaled maximum ratio (adjusted by subtracting the corresponding average values in the background data taken just before the vehicle profile) for each of several profiles of eight common NORMs. The number of profiles for each NORM category is given in parentheses in the figure legend.

3. DISTINGUISHING NORM TYPES

From knowledge of the key gamma emitters in the constituents of common NORM types, the ratio of low-energy to total-energy gamma counts is anticipated to be approximately the same for NORM as for the background, whereas for most threats, this ratio is anticipated to increase. Therefore, profile counts are normalized using the ratio

$R_L = (P_L - B_L)/\sqrt{B_L}$,

where B_L is the background low-energy count taken just before the vehicle is sensed and P_L is the profile low-energy count. Burr and Myers [4] focused on 2-window examples; other systems in use have multiple windows (up to 8 energy windows) and still other systems are expected to have 128 or more windows. Data mining described in Burr and Myers [4] illustrated that no increase is observed in R_L because of the types of cargo evaluated. Therefore, an alarm strategy that alarms if the gamma counts exceed a threshold *or* the gamma ratio exceeds a threshold could potentially lower the nuisance alarm rate on NORM cargo. This scenario would be the case in a system monitoring both counts and count ratios that revised the gamma counts alarm threshold upward to compensate for having added another alarm rule involving the ratio. This strategy makes two key assumptions. First, most NORM must have approximately the same low-energy-to-total-energy count ratio as background. Second, all or most threats should emit more low-energy than high-energy gammas.

Regarding the first assumption, Ely *et al.* [5] report initial empirical results that are promising. Most of the examined NORM vehicles that caused count alarms would not have alarmed based on a gamma ratio criterion. However, unless the gamma counts alarm threshold were raised to compensate for adding the ratio alarm rule, these NORM vehicles would still alarm on the gamma counts criterion. It is hoped that results can be further improved by using more than two energy windows in the future.Regarding the second assumption, not all threats emit more low- than high-energy gammas. It is therefore possible that the revised-upward count criterion would lead to a lower detection probability for some threats. Some type of cost-benefit analysis using assumed probabilities of each threat type could be used to formally justify the inclusion of the ratio criterion. Detailed studies of threat signals are in progress (but are not available for public release), where threats are defined on the basis of the source and shielding. Also, various "masking scenarios" (involving loads having both low- and high-energy gamma emitters) could arise. Without more detailed descriptions of which NORM types do not emit preferentially at low energies and which threats do, it is not possible to quantify the possible performance gain by including the ratio criterion. Therefore, the potential merit of the ratio criterion was considered in Burr *et al.* [6] by evaluating currently available ratio data with and without injected signals. These injected signals were synthetic counts added to real vehicle profiles. The next generation of RPMs will include better spectroscopic resolution, enabling other NORM-discrimination strategies; in the meantime, it is useful to characterize the currently available gamma counts and count ratios. Burr and Myers [4] illustrated only moderate success at distinguishing among common NORMS. Fortunately, Figure 3 illustrates qualitatively that the measured high- and low-energy percentages can distinguish common NORMs from two threat isotopes [highly enriched uranium (HEU) and weapons-grade plutonium (WGPu)]. Figure 3(a) uses one real example from each of five common NORMs and from HEU and WGPu. Figure 3(b) uses multiple real examples of each isotope.

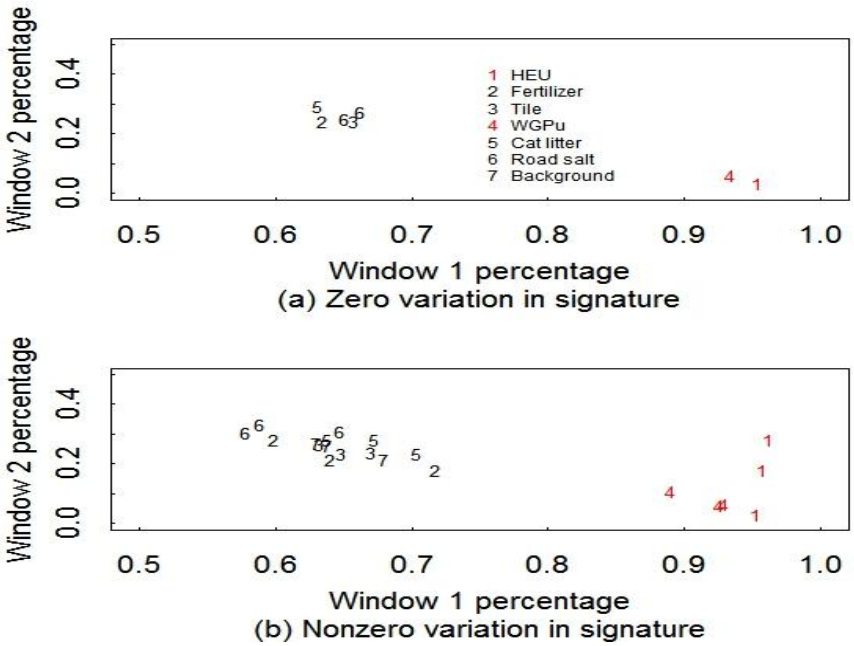

Figure 3. Window 2 (high energy) versus window 1 (low energy) for five NORM types and two threat types (HEU and WGPu) with and without variation in the signatures.

4. AVAILABLE DATA

Figure 4 plots counts in energy bin 1 (low energy) and energy bin 2 (high energy) in each of two real non-alarming profiles. Notice that background suppression leads to slightly lower counts as the vehicle approaches the detector panels. Plots of the corresponding R_L did not alarm and demonstrate that suppression of R_L is negligible. Although we have access to hundreds of thousands of profiles and tens of thousands of alarming profiles, relatively few alarming profiles have a description of the cargo available. Burr and Myers [4] selected 406 alarming profiles (all alarms were gamma alarms) that had electronically available cargo labels; of these, 235 were selected as representing the 8 most frequent NORMs.

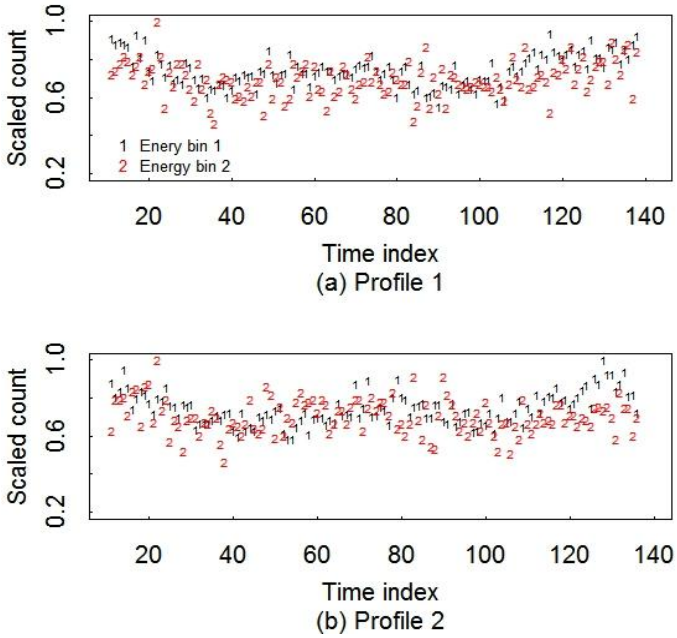

Figure 4. Counts in energy bin 1 (low energy) and energy bin 2 (high energy) in two real profiles.

5. FEATURES FOR PATTERN RECOGNITION

5.1 Feature Extraction

The 46 features extracted from each of the 4 panels are described in Burr and Myers [4]. For example, features 1 to 4 are the maxima of the low gammas, high gammas, neutrons, and gamma ratio, respectively, and features 5 to 8 are the corresponding average counts over the profile. A total of 184 (46 *x* 4) features were extracted.Features 9 to 40 from Panel 1 are correlations between the four profiles (low gammas, high gammas, neutrons, and gamma ratio), with the corresponding four example profiles from each of the eight most common NORM types: abrasives, aluminum oxide, potassium chloride, tableware, cement, potash, bentofix, and cat litter. Features 9 to 12 are correlations with the abrasives example profile; Features 13 to 16 are correlations with the aluminum oxide example profile, etc. Note that using the correlation is one way to compare peak magnitudes and shapes. The correlation was computed by first aligning each profile to the same length (we used length 150) and then computing the usual correlation between profiles *x* and *y* as

$$r_{xy} = \sum_{i=1}^{150}(x_i - \bar{x})(y_i - \bar{y}) / \{\sum_{i=1}^{150}(x_i - \bar{x})^2 \sum_{i=1}^{150}(y_i - \bar{y})^2\}^{1/2}.$$

The example profile for each of the eight classes was randomly chosen among the available examples.Features 41 and 42 are the peak width at half the maximum amplitude for low- and high-energy gamma counts, respectively. Features 43 and 44 are the average neutron count (in standard deviation units above the background) during the gamma low-energy peak and not during the gamma low-energy peak. Features 45 and 46 are the same as 43 and 44, but for the gamma high-energy peak.

5.2 Feature Selection

Section 4.1 described 184 features extracted from each profile, but as expected, not all 184 features were found to be effective. An important issue that impacts the misclassification rate of any classifier is the choice of features to be used to make classification decisions. This choice is called feature selection, or sometimes model or predictor selection, depending on the context. Roughly, the well-known "curse of dimensionality" [9] suggests that the performance of most classifiers is improved by using only those features that are somehow related to the response and having a "very large" number of training observations per feature. Exactly what qualifies as very large has not been rigorously defined; however, a paper by Silverman [10] contains approximate results that relate to the issue of "what is large" [11].

One common method to select the best features from many candidates is to evaluate the between-group variance σ_B^2 and the within-group variance σ_W^2 of each feature. Features having a large ratio σ_B^2 / σ_W^2 are good candidates for discriminating among groups. In our context, a group is a collection of profiles having the same cargo label.A strategy that is often effective for displaying multiple features involves multidimensional scaling (MDS) [12]. A common application of MDS is to first use as many features as desired to compute the pairwise distances among all observations. Next, MDS is used to approximate this matrix of pairwise distances. MDS is similar to principal components analysis but has the goal of finding a few (Burr and Myers [4] used two) coordinates to closely approximate the matrix of pairwise distances among observations. The two MDS coordinates could then be used to closely approximate the original matrix of pairwise distances and can effectively convey the extent of separation among various groups in the data in a simple two-dimensional plot.Burr and Myers [4] applied MDS to all of the features, to features from only one sensor panel at a time, or only to features that have large σ_B^2 / σ_W^2. One aspect of MDS is that results are sensitive to how the features are scaled, and it is prudent to empirically evaluate effectiveness using the original scale for each feature and also to rescale each feature to have the same variance. When features have widely differing variances, there is a tendency for high-variance features to carry most of the weight in defining the MDS coordinates. This tendency can be desirable or not, depending on the context. We have observed no clear improvement in group separations as more features are included in defining the distances upon which MDS operates to produce two coordinates. However, adding more features can reduce the misclassification rates, suggesting that applying pattern recognition methods to more than two MDS coordinates will result in lower misclassification rates. For example, we show below that using three MDS coordinates does in this case lead to the lowest misclassification rates.

6. PATTERN RECOGNITION METHODS

Pattern recognition is typically defined as the task of using predictor variables (such as low-energy gammas) to predict a categorical response variable (such as what type of NORM is present in an alarming vehicle). In the typical pattern recognition problem, the data consist of n cases of (C, X) pairs, where the integer $C \in (1, 2,..., J)$ is the class and X is a p-dimensional predictor vector. The goal to use X to predict the class C is also sometimes called a classification problem or a supervised learning problem. Regarding notation, matrices, vectors, and scalars can be distinguished by context and definition. For example, C is a scalar and X is a p-dimensional vector. Predictor variables are sometimes called features; we use the terms "predictors" and "features" interchangeably.There are many traditional and modern approaches to DA. Some of these approaches attempt to estimate the probability density of a predictor vector, X, given its class [i.e., the conditional probability $P(X|C)$] by assuming some convenient distribution for $X|C$. In our case, the classes are the various types of NORM, and the predictors X are the features described in Section 4. We consider methods that assume that the distribution is stationary over time. We include nonparametric methods (which make essentially no assumptions regarding the form of the underlying probability distributions), such as nearest neighbor classifiers, flexible discriminant analysis, distribution-free density-estimation-based DA (DFDA), and learning vector quantization [9], [12]. We also include parametric methods, such as linear discriminant analysis. If the class conditional probabilities are not stationary over time, then methods described in a paper by Black and Hickey [13] could be considered.In our experience, several approaches should be compared using held-out test data and/or cross validation [9], [12] to provide performance estimates for each approach. We next briefly describe each technique used (see papers by Hastie et al.[9] and Venables and Ripley [12] for details of these techniques). Several of these methods [K-nearest neighbor (K-NN), a type of neural network, and linear discriminant analysis (LDA), a method based on principal components] were described and successfully applied by Dravogiv and Onjia [14] to classify soil samples.

6.1 Linear Discriminant Analysis (LDA), Mixture DA (MDA), and Flexible DA (FDA)

LDA is the original pattern recognition technique created by R. Fisher in the 1930s; it has a linear decision boundary separating classes. LDA assumes that the data for each class arise from a multivariate normal distribution, with the classes differing only in the mean vector.MDA is a natural extension of LDA, which allows any number of mean vectors, instead of just one, for a given class.FDA attempts to exploit the fact that LDA can be derived by repeated linear regression of the class (viewed as a response) on the predictors. The linearity assumption then is dropped, and any nonlinear regression methods (including neural networks) are allowed. Its construction of nonlinear class boundaries has been shown to behave similarly to that of support vector machines, which have become relatively popular [9].

6.2 Distribution-Free Discriminant Analysis (DFDA)

This nonparametric ("distribution-free") method relies on kernel-based density estimation (an improved version of a "histogram" density estimate) to estimate the class probability $P(X|C)$ for each class [7].

6.3 K-Nearest Neighbor (K-NN) Methods

These techniques classify a given case in the testing set according to the classes assigned to the nearest (in the predictor space) k cases in the training set by using majority rule. Burr and Myers [4] used the values of 1, 3, 5, and 10 for k.

6.4 Neural Network: Learning Vector Quantization (LVQ)

The LVQ method finds a modest number of representative vectors (having the same dimension as each input vector) per class. The class of a given case is predicted by applying the K-NN method to these representative vectors rather than to the original data. During training, the representative vectors are adjusted to improve performance on the training data.

6.5 Classification and Regression Trees [Recursive Partitioning (RPART)]

This technique uses a decision tree. Freeware versions are available, such as RPART in the statistical programming language R (http://www.r-project.org). Classification trees, which learn effective rules for partitioning the data on the basis of "high/low" thresholds (such as "is predictor 1 less than threshold 1"), have proven to be competitive on many real data sets.

7. PATTERN RECOGNITION RESULTS

Burr and Myers [4] considered three cases. In the first, the goal was to recognize each of eight NORM materials. In the second, the goal was to recognize cat litter. Similarly, in the third case, the goal was to recognize potash from non-potash. All pattern recognition methods described in Section 5 were implemented in the R statistical programming language (http://www.r-project.org/). The average misclassification rate was based on partitioning the respective data set into 67% for training and 33% for testing and recording the resulting average misclassification rate over 100 random repeats of the partitioning. All reported percentages were repeatable to within ±1%. For the "eight classes" case, no method performed very well. However, using 46 or 184 predictors is generally better than using only 4 predictors. And using MDS with three coordinates is the overall best choice, resulting in nearly all methods having approximately a 28% to 32% misclassification rate. For the "cat litter or not" case, using LDA with 46 predictors from 1 panel or DFDA with 46 predictors from 4 panels performed well, with approximately a 17% misclassification rate. But again, using MDS with three coordinates was the best choice, resulting in approximately a 13% misclassification rate by LVQ. For the "potash or not" case, using 4 or 46 predictors and 4 panels with any of several methods results in an approximately 8% misclassification rate, which is surprisingly better than the two 1-panel results or the 184 predictors and 4-panel result. We have not found the reason for this discrepancy; however, by inspecting figures such as Figure 3, we expect to recognize potash from non-potash fairly well. Again, using MDS with two or three coordinates resulted in the lowest misclassification rates for the various methods (6% to 10%).

8. OTHER DATA MINING TASKS

The two other challenges mentioned in the Introduction section that data mining can help address are (1) the impact of drifting neutron and gamma background rates and (2) RIID performance. Drifting neutron background rates was studied in Burr and Hamada [15] by analyzing many thousands of real neutron count rates. Such data mining informed simulation models to evaluate the impact of different averaging time periods to balance the need for large-enough averaging times to reduce random estimation error and small-enough averaging times to mitigate the effect of the true neutron background rate varying over time. RIID performance continues to be evaluated in ongoing laboratory and field experiments. Burr and Hamada [16] illustrate some of the difficulties with RIID performance using NaI detectors (which have higher energy resolution than detectors that are currently fielded in primary screening). The ASP program mentioned in the Background section currently has been cancelled because of the poor performance of fielded NaI RIID algorithms. However, various agencies within and outside the US continue to evaluate and improve NaI-based RIID algorithms [16].

9. DISCUSSION AND SUMMARY

The misclassification rates for recognizing common NORMS are high; it is currently unknown whether they are too high to enable an effective alarm resolution strategy. One goal is to have high confidence in a timely manner that alarming NORM cargo is in fact innocent NORM. Current efforts to improve the ability to recognize common

NORM involve using more energy windows in the primary screening. Results from these multi-window options can be compared with results such as those presented here to quantify the benefit of using more than two energy windows. Although formal studies have not yet been published, we can reasonably expect that similar experiments to distinguish NORMs from threat isotopes will have lower misclassification rates. This paper has reviewed data mining primarily for gamma detectors in the role of recognizing commons NORMs. A short description of data mining for the impact of drifting neutron and gamma background rates and for RIID performance was also given.

10. REFERENCES

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

[2]. R.T. Kouzes, J.H. Ely, R.R. Geelhood, E.A. Lepel, J.E. Schweppe, D.J. Siciliano, D.J. Strom, and R.A. Warner, Naturally occurring radioactive materials and medical isotopes at border crossings, IEEE Nuclear Science Symposium—Conference Record, 1448–1452 (2004).

[3]. R.T. Kouzes, J.H. Ely, J. Evans, W. Hensley, E. Lepel, J. McDonald, J. Schweppe, E. Siciliano, D. Strom, M. Woodring, Packaging naturally occurring radioactive materials in cargo at U.S. borders, *Transport, Storage & Security of Radioactive Material* **17**(1), 11–17 (2006).

[4]. T. Burr, K. Myers, Signatures for several types of naturally occurring radioactive materials, *Applied Radiation and Isotopes* **66**, 1250–1261 (2008).

[5]. 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 (2), 373–387 (2005).

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

[7]. T. Burr, J. Doak, Distribution free discriminant analysis, *Intelligent Data Analysis* **11**, 651–662 (2007).

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

[9]. T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning," Springer: New York (2001).

[10]. B.W. Silverman, "Density Estimation for Statistics and Data Analysis," 1st ed, Chapman and Hall, London (1986).

[11]. A. Ghosh, P. Chaudhuri, D. Sengupta, Classification using kernel density estimates: multiscale analysis and visualization, *Technometrics* **48**(1), 120–132 (2006).

[12]. W. Venables, B. Ripley, "Modern Applied Statistics with Splus," 3rd ed., Springer: New York (1999).

[13]. M. Black, R., Hickey, "Maintaining the performance of a learned classified under concept drift," *Intelligent Data Analysis* **3**(6), 453-477 (1999).

[14]. S. Dragovic, A. Onjia, Classification of soil samples according to geographic origin using gamma-ray spectrometry and pattern recognition methods, *Applied Radiation and Isotopes* **65**(2), 218–224 (2007).

[15]. T. Burr, M.S. Hamada, Data analysis in support of radiation portal monitoring, *International Journal of Research and Reviews in Applied Science* **14**(1), 1–16 (2012).

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

Funding acknowledgment: Department of Homeland Security