Anomaly Detection in Radiation Sensor Data With Application to Transportation Security

Olufemi A. Omitaomu, *Member, IEEE*, Auroop R. Ganguly, *Member, IEEE*,
Bruce W. Patton, and Vladimir A. Protopopescu

***Abstract*—In this paper, we present a new approach for detecting trucks transporting illicit radioactive materials using radiation data. The approach is motivated by the high number of false alarms that typically results when using radiation portal monitors. Our approach is a three-stage anomaly detection process that consists of transforming the radiation sensor data into wavelet coefficients, representing the transformed data in binary form, and detecting anomalies among data sets using a proximity-based method. The approach is evaluated using simulated radiation data, and the results are encouraging. From a transportation security perspective, our results indicate that the concomitant use of gross count and spectroscopy radiation data improves identification of trucks transporting illicit radioactive materials. The results also suggest that the use of additional heterogeneous data with radiation data may enhance the reliability of the detection process. Further testing with real radiation data and mixture of cargo is needed to fully validate the results.**

***Index Terms*—Anomaly detection, border and transportation security, illicit radioactive material, weigh station.**

I. Introduction

THE transportation of illicit radioactive materials that have evaded security checks at ports or border crossings poses a security risk to highly sensitive and secure locations. As an illustration, a maritime container that has evaded security checks at the Port of Mobile in Alabama and transferred to a commercial truck could be in any of the cites in the southeast region of U.S. within 12 h of the highway travel time and any of the cities in the entire eastern half of the United States within 24 h. Therefore, the ability to detect the transportation of illegitimate radioactive material on U.S. highways has become an important national security research objective [1]. Illicit radioactive materials are special nuclear materials (SNMs) that can be used for making "dirty bombs." Some of these materials are readily available as medical isotopes in hospitals. The detection processes are currently being performed with radiation portal monitors (RPMs) at border crossings, ports, and weigh station test beds.

When a truck triggers an alarm (i.e., identified as a possible security risk) using the RPM at a weigh station test bed, the inspection officer requests that the truck be taken through a secondary inspection process and, possibly, a tertiary inspection process. The additional inspection processes involve the collection of some supplementary data such as spectroscopy data for further analyses and, possibly, manual inspection of the truck. These additional processes cause truck delays and increase the operating costs of the weigh stations. A flowchart representation of the inspection processes is depicted on the lower half of Fig. 1.

Generally, the initial alarm is a false alarm, and this has been attributed to several factors. First, there is some legitimate cargo, such as kitty litter and fertilizer in commerce and medical isotopes in hospitals, that contain naturally occurring radioactive materials (NORMs) or technologically enhanced naturally occurring radioactive materials (TENORMs), respectively. Both will trigger alarms when using the RPM [2]. These alarms are triggered by radioactive materials, which are generally not the target, and constitute nuisance alarms. Second, the RPM uses a plastic scintillator material because it has relatively good sensitivity and low cost compared with other detection materials, such as doped sodium iodide or germanium [3]. This makes the RPM only suitable for analyzing gross count data. However, when illicit radioactive materials are mixed with legitimate NORMs or TENORMs in commerce or shielded with physical objects such as thick lead metal boxes, the gross count data may not be sufficient for detecting the shielded materials. Other problems associated with the use of RPMs for detecting illicit radioactive materials in trucks have been investigated [2]–[7]. The overall conclusion of these studies is that there is a need for better techniques that could minimize, if not eliminate, the number of false alarms and be able to detect shielded illicit radioactive materials. The use of additional data such as spectroscopy data during the secondary or tertiary inspection processes does not solve the initial problem of reducing the number of false alarms because the spectroscopy data are used *after* the initial false alarm. Furthermore, the analyses of the spectroscopy data require contacting the reachback support team at designated locations far away from the weigh stations. This further increases the costs of operating weigh stations.

One approach to reduce the number of false alarms during the primary inspection process and eliminate (or drastically reduce)

Manuscript received September 19, 2007; revised August 6, 2008 and October 22, 2008. First published April 28, 2009; current version published June 2, 2009. This work was supported by the Laboratory Directed Research and Development Program of the Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract DE-AC05-00OR22725. The Associate Editor for this paper was D. Zeng.

O. A. Omitaomu, A. R. Ganguly, and V. A. Protopopescu are with the Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA (e-mail: omitaomuoa@ornl.gov; gangulyar@ornl.gov; protopopesva@ornl.gov).

B. W. Patton is with the Nuclear Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831 USA (e-mail: pattonbw@ornl.gov).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2009.2020190

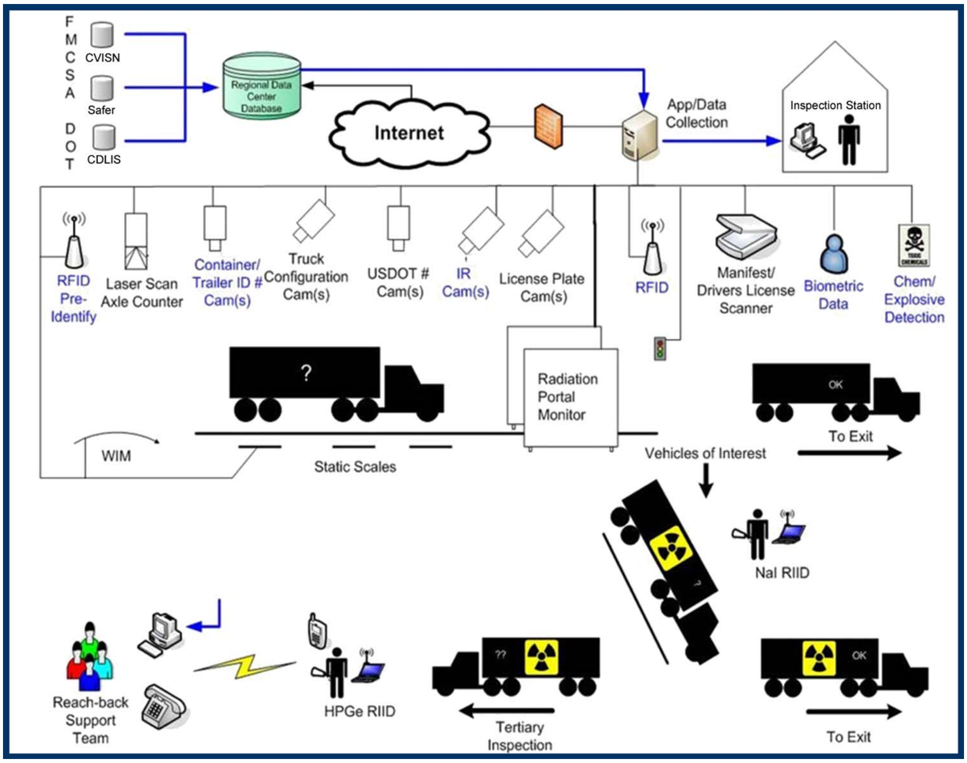

Fig. 1. Truck inspection process flow diagram at a weigh station test bed.

unnecessary truck delays, as well as the costs associated with such delays, is to use multiple sensor data such as radiation and truck weight-related data during the primary inspection process. Some of the other possible sources of additional data are depicted in the upper half of Fig. 1. Such a decision, however, requires proven techniques for analyzing each of the sensor data. Furthermore, the techniques must be able to reduce (if not eliminate) the number of false alarms and eliminate missed detections. In addition, the approach should provide outcomes that would aid the decisions of the inspection officers within few seconds because commerce cannot be interrupted or delayed beyond the normally allowed time to process a truck. These requirements motivate the approach proposed in this paper. The proposed approach is intended for each of the radiation (spectroscopy and gross count) data. Therefore, the focus of this paper is the problem of detecting the transportation of illegitimate radioactive materials in trucks using simulated radiation data with the assumption that such materials may be mixed with legitimate NORMs or TENORMs in commerce or shielded with a thick lead metal box.

Radiation data provide isotopic identification of radioactive materials. The gross count (radiation) data measure the total radiation counts in a material, whereas the spectroscopy (radiation) data indicate the isotopes that are present in the material. Radiation data have some characteristics that make them a challenge to use general-purpose anomaly detection approaches. These characteristics include subtle differences among data of different NORMs that may be difficult to quantify in the data domain and peak information that is useful for discriminant analysis. Furthermore, the relative rarity of security-related violations makes the problem even more challenging. To address these challenges, we propose a three-stage anomaly detection approach that consists of techniques that, independently, have extensively been studied in other domains and used in several applications. However, to our knowledge, this is the first attempt at combining these techniques for anomaly detection applications. The three stages are summarized in the list that follows.

1) *Data transformation.* The radiation data are transformed into the wavelet domain for effective discrimination while retaining the sequence of the data in time.
2) *Clipping wavelet coefficients (WCs).* The WCs are represented in binary form to distinguish small peaks from large peaks.
3) *Proximity-based anomaly detection method.* Radiation binary data are identified as anomalous if their *local proximity range* exceeds the *global proximity range* (GPR) for the given (or baseline) data. The GPR is defined as $\text{MLP} + (Z \times \text{SLP})$, where MLP is the *mean of all local proximity ranges*, SLP is the *standard deviation of all local proximity ranges*, and Z is called the *proximity factor* (threshold) set by the user.

The remainder of this paper is organized as follows: In Section II, we describe the process for simulating the radiation data used in this paper. The proposed anomaly detection approach is described in Section III. The results of the application of our approach to transportation security using 40 simulated radiation data in six different scenarios are discussed in Section IV. Some of the implications of our results for transportation security are presented in Section V, along with some future areas of research.

II. Simulation of Radiation Data

To test our approach, labeled sets of data representing both normal and anomalous trucks (cases) for different relevant scenarios are needed. Since trucks transporting illicit radioactive materials are rare, obtaining data sets of anomalous trucks is a challenge. One approach to address this challenge is to collect large amount of data under a controlled environment, in which different loading patterns and shielding agents are used with different combinations of NORMs, TENORMs, and SNMs to describe different possible scenarios. However, such data collection exercise could be very expensive, and only known or conceivable scenarios could be considered. A less expensive approach is to develop simulated data that would mimic the real-world data for initial investigations and assessments of the proposed approach. The simulated data could help us answer two fundamental questions: 1) Can we get additional insights from the use of only the spectroscopy data if the SNMs are mixed with legitimate cargo in commerce? 2) Do we need other data besides gross count and spectroscopy data? It could also provide information that will guide the design and implementation of controlled experiments with real-world data. The approach for simulating the radiation data used in this paper is described in this section.

A. Simulation Process

The process of collecting radiation data is modeled using the Monte Carlo N-particle transport code, i.e., MCNP5 [8]. The use of MCNP5 allows the simulation of both radioactive NORM materials and possible illicit radioactive materials, if desired. The MCNP5 is a general-purpose code that can be applied to several applications, including radiation shielding, detector design and analysis, and fission and fusion reactor design. It can be used in several transport modes, such as neutron only, photon only, and combined neutron/photon/electron transport. The neutron energy regime is from 10^{-11} to 20 MeV for all isotopes and up to 150 MeV for some isotopes; the photon energy ranges from 1 keV to 100 GeV; and the electron energy range is from 1 keV to 1 GeV. The MCNP5 simulations of the data used in this paper were performed on the 762 CPU Intel Xeon 3.0-GHz cluster configuration service daemon located at the Oak Ridge National Laboratory (ORNL), Oak Ridge, TN. Typically, eight CPUs were utilized for each calculation, with total CPU run times ranging from 8 to 12 h. This necessitates the use of parallel computers to reduce the time required to complete each calculation. The high-performance computing facilities at ORNL were then used and reduced the computation time by more than 60%.

An 18-wheel truck is modeled as an aluminum shell of the appropriate interior dimensions and shell thickness of typical commercial trucks. A schematic of the simulation setup consisting of the truck, the RPM, and the pavement of a typical highway is shown in Fig. 2. It should be noted that for the purposes of this simulation, the wheels and the cab of the truck were neglected, as their inclusion in the model would have a little effect on the simulation results other than to increase the run time. The shell is set up to contain a full or partial load of the desired NORM material. Kitty litter is our choice of NORM in these simulations for various reasons: 1) It is generally recognized as one of the NORMs that cause the most false alarms at weigh stations; 2) it is easy to procure; and 3) its high background level makes it a likely material for shielding illicit radioactive materials during transportation.

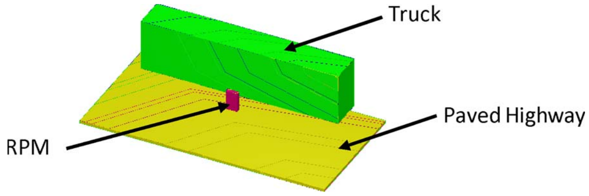

Fig. 2. Schematic of the simulation process.

The experimental gamma spectrum of kitty litter provided by one of the manufacturers of RPMs was input into the MCNP5 via the source definition capability, with the energy bins for gamma emission corresponding to those of the RPMs. As a check, the MCNP5 detector energy spectrum response was compared with an experimental response from a weigh station RPM for a truck load of kitty litter; the MCNP5 was able to quite accurately reproduce the energy spectrum, given a sufficient number of particles to obtain good statistics.

During each simulation, the truck was moved from −1000 to +1000 cm in 50-cm intervals. At each location, a computation of the detector response to a truck containing some NORMs was performed. Typically, 100 million histories were run at each truck location to get reasonable statistics in the output data. The scenarios used for generating the simulated data are based on two major challenges with the current detection process. A total of 11 different scenarios were simulated, but only six scenarios are presented and discussed in this paper. Each simulated scenario consists of either five or seven cases (each case represents a truck in the real world), whereas each case has one or two data types (spectroscopy data and/or gross count data), and each data type is the average of ten simulations.

B. Simulation of Hidden Isotopes in Radiation Data

One of the challenges of the current process is the ability to detect an illicit SNM source shielded with legitimate NORMs. To address this challenge, five different simulated cases were generated using kitty litter as the NORM source; this simulation is called Scenario I in Table I. Only spectroscopy data are simulated for each of the five cases in this scenario. Gross count data are not considered in this scenario based on the prior knowledge that a mixture of radioactive materials is better captured by the spectroscopy data. The simulation is set up in such a way that three of the five cases contain only the NORM and two cases contain both the NORM and one SNM source (Cs-137 or Ba-133), respectively; the NORM in these two cases serves as a shield for the SNM. The three cases are examples of normal cases, whereas the two cases are examples of anomalous cases. The two selected SNM sources are some of the readily available medical isotopes found in commerce. The NORMs used in the three normal cases are modeled as

TABLE I
COMPOSITION OF SIMULATED RADIATION DATA

Scenario	Case	Data Types	
		Gross Count Data	Spectroscopy Data
I	1	-	100% NORM + SNM (Cs-137)
	2	-	100% NORM (different seeds)
	3	-	100% NORM (different seeds)
	4	-	100% NORM + SNM (Ba-133)
	5	-	100% NORM (different seeds)
II	1	50% NORM	50% NORM
	2	10% NORM	10% NORM
	3	70% NORM	70% NORM
	4	30% NORM	30% NORM
	5	60% NORM	60% NORM
	6	40% NORM	40% NORM
	7	20% NORM	20% NORM
III	1	50% NORM	50% NORM
	2	10% NORM	10% NORM
	3	70% NORM	70% NORM
	4	30% NORM	30% NORM
	5	60% NORM	60% NORM
	6	40% NORM	40% NORM
	7	50% NORM + box	50% NORM + box
IV	1	50% NORM + box	50% NORM + box
	2	10% NORM + box	10% NORM + box
	3	70% NORM + box	70% NORM + box
	4	30% NORM + box	30% NORM + box
	5	60% NORM + box	60% NORM + box
	6	40% NORM + box	40% NORM + box
	7	50% NORM + box	50% NORM + box
V	1	30% NORM + box	30% NORM + box
	2	10% NORM	10% NORM
	3	30% NORM + box	30% NORM + box
	4	20% NORM	20% NORM
	5	30% NORM	30% NORM
	6	20% NORM + box	20% NORM + box
	7	50% NORM	50% NORM
VI	1	50% NORM	50% NORM
	2	50% NORM + box	50% NORM + box
	3	70% NORM + box	70% NORM + box
	4	70% NORM	70% NORM
	5	60% NORM	60% NORM
	6	40% NORM	40% NORM
	7	50% NORM + box	50% NORM + box

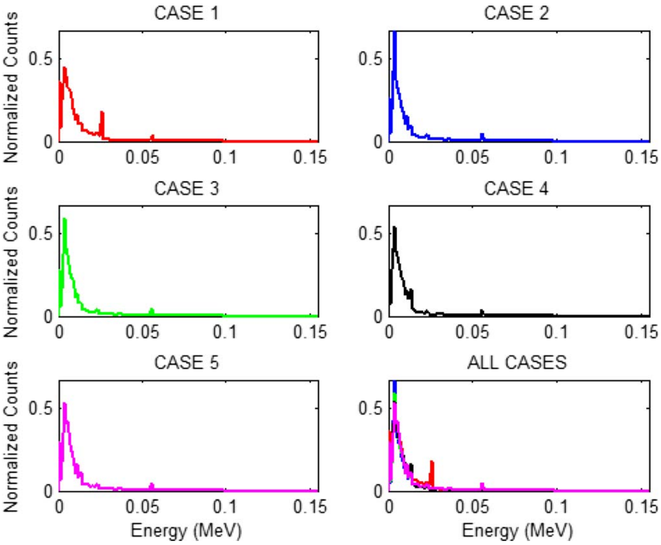

Fig. 3. Five cases in Scenario I. Cases 2, 3, and 5 contain only the NORM source, whereas Cases 1 and 4 contain the NORM and an SNM source.

NORMs from different geographical regions by using different seeds. Although this disparity exists among these three cases as in the real world, these cases should be identified as examples of normal cases. This is necessary since various cargo found in commerce are from different geographical locations, countries, and manufacturing facilities, and the little differences in their compositions should not be drawbacks for the detection process. For each case, the aluminum shell is set up to contain a full load of the desired material(s). The five cases are plotted in Fig. 3. The plots in this figure reveal that the five cases are very similar, and it is difficult, even visually, to tell the differences, particularly among Cases 2, 3, 4, and 5. The only case that looks very different from the rest is Case 1, because it has an extra peak at 0.025 MeV. This is a challenge because it is very difficult to establish a baseline for normal cases since Case 4 is very similar to Cases 2, 3, and 5. The results of the application of our approach to these data sets are described in Section IV-A.

C. Simulation of the Shielded SNM and Low Percentage of the NORM

Another challenge of the current detection process is that with sufficient shielding, detection would be almost impossible for even the most sensitive detector, thereby resulting in missed detections. We address this challenge using a series of simulated scenarios (Scenarios II–VI in Table I). In these scenarios, we consider the possibility of transporting different percentages of the NORM source with or without shielding with a physical object such as a thick lead metal box. The objective is to be able to detect cases with a metal box, regardless of the percentage of the NORM. Each of the five scenarios consists of seven cases, and each case has two data types (spectroscopy and gross count data).

For these scenarios, we again used kitty litter as the NORM source. For each case, the aluminum shell is set up to contain a fraction (expressed as a percentage) of the desired material to fool the detection approach. In addition, some cases have a thick empty metal box placed with the NORM source. This empty metal box serves as the “idealized” shielding agent for the SNM. No SNM source was placed in the metal box because the assumption for this setup is that the box walls would be so thick as to preclude the detection of any radiation source that may be inside the box. Therefore, if any approach can detect the presence of an empty metal box, it should easily be able to detect a box containing an SNM source. The compositions of these five scenarios (Scenarios II–VI) are shown in Table I.

Scenario II will help determine if the approach can detect any difference in the cases with different percentages of the NORM. Scenarios III–VI will help determine if the cases with a box can be distinguished from the other cases. The other five

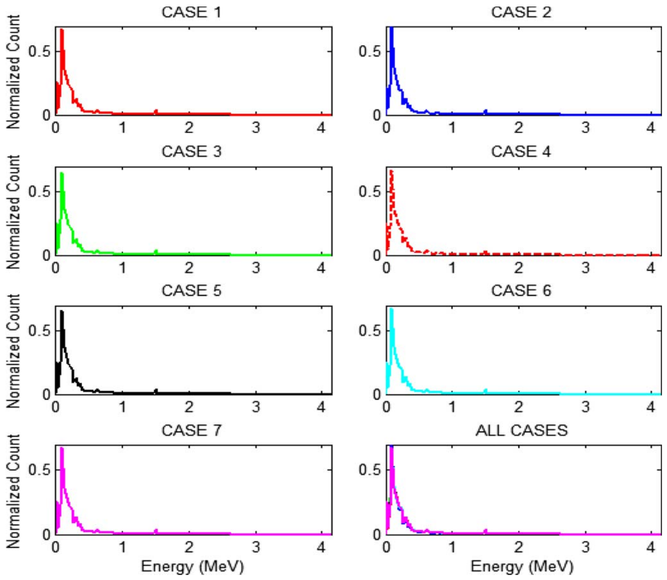

Fig. 4. Spectroscopy data for the seven cases in Scenario III.

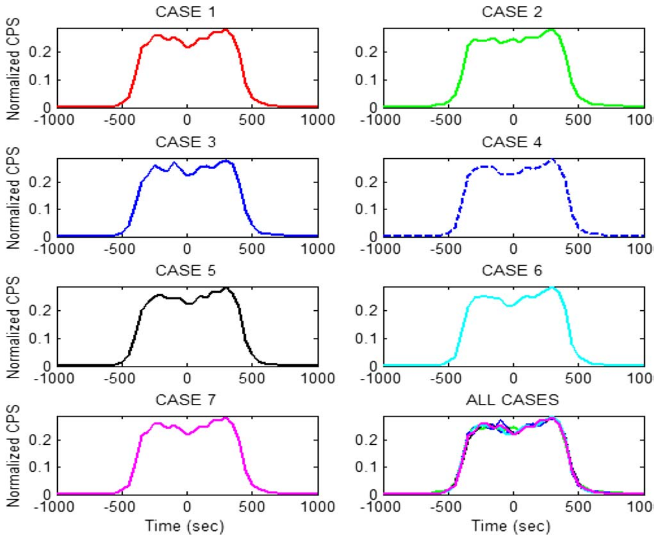

Fig. 5. Gross count data for the seven cases in Scenario IV.

scenarios (which are not discussed in this paper) share similarities with these five scenarios. The use of two data types in these five scenarios will help determine which of the two data types is useful for quantifying changes in the percentage of the NORM and/or detecting the presence of a box.

The plots of some of these scenarios are shown in Figs. 4 and 5. The plots of the spectroscopy data from Scenario III (Fig. 4) do not show any indication that a metal box is present in Case 7 or that a different percentage of the NORM is used in any of the cases. This is also true if we consider the gross data in Scenario IV, as shown in Fig. 5. This further confirms the fact that the differences among cases are usually very small and are not obvious in the original domain. These small differences are a challenge for any detector, as well as any anomaly detection approach, because it is difficult to establish a single baseline. The results of the application of our approach to these scenarios are described in Section IV-B.

III. Anomaly Detection Approach

In this section, we present a new approach for detecting illicit radioactive materials in trucks using radiation data. Our approach is based on the assumption that most of the given sets of data represent normal cases and are, in a certain sense, similar and that only a few data sets are "abnormal." The approach consists of three stages, as depicted in Fig. 6. In the following sections, we describe each of the stages and provide the framework for their implementations.

A. *Data Preprocessing Based on Wavelet Transformation*

One of the most difficult challenges in detecting anomalies from radiation data sets is that the smallness of the differences between signals in the data domain may obfuscate discrimination. Again, consider the five cases in Scenario I, as shown in Fig. 3.

It is clear from those plots that the difference between the five cases is subtle. To address this challenge, it is often advantageous to transform the data sets onto another domain, where the minute differences are magnified. For this application, the choice of the new domain and the corresponding transformation are determined by the following considerations.

The temporal sequence of various peaks in the data sets is an important feature here, since peaks are indicators either of the isotopes that generate the data or of radiation counts. Therefore, for effective similarity matching, the selected signal processing technique should be able to retain or retrieve the temporal sequence of the data points. Several signal processing techniques such as the Fourier transform (FT) method [9], the empirical mode decomposition (EMD) method [10], the Savitzky–Golay (SG) smoothing filter [11], and the wavelet transform (WT) method [12] have routinely been used for signal analysis and anomaly detection. Among these techniques, FT does not retain the time dependence of the data; moreover, FT makes physical sense only if the original signal is stationary, which, in general, is not the case. Among the remaining three techniques, SG tends to smooth the signals, which results in loss of information. Thus, EMD and WT remain as possible candidates for the signal transformation.

The EMD method decomposes the original signal into *several* empirical modes, whereas the WT method transforms it into a *single* set of WCs. In this paper, we choose to transform the original data into WCs for anomaly detection purposes because it is computationally efficient to work with a single set of WCs than several empirical modes. The idea of detecting anomalies in the wavelet domain has also been explored in other applications [13]–[15]. For a good introduction to WTs, see [16] and [17].

Two of the five cases in Fig. 3 are used as an illustration in Fig. 7. The top plots are the original data, and the middle plots are the respective WCs of the two original data. We can see that although there seems to be no difference between the two data in the original plots, the plots of the WCs indicate some differences in the data. These differences are even more distinct in the binary domain, as shown in the bottom plots in Fig. 7. The implication of the WT stage is further examined in Section IV-A. It is shown later (see Sections IV-A and B)

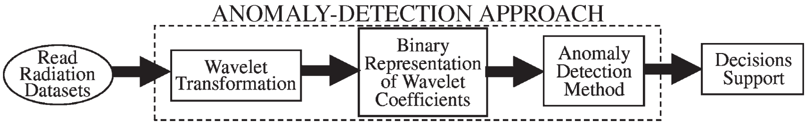

Fig. 6. Block diagram for the proposed anomaly-detection approach.

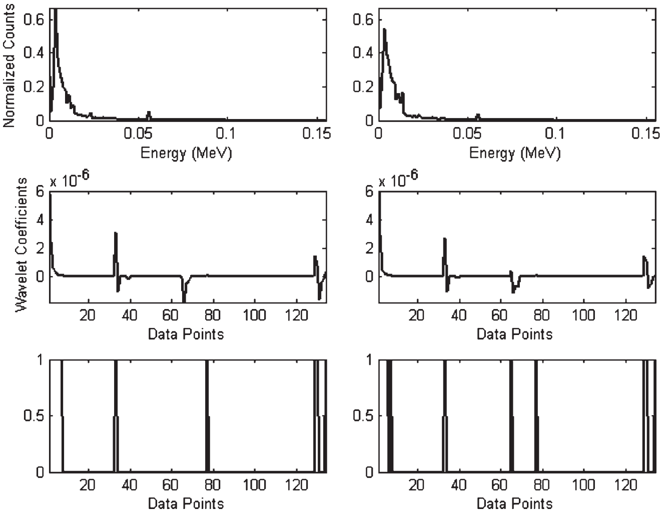

Fig. 7. (Top plots) Two of the five cases in Scenario I, (middle plots) their WCs, and (bottom plots) binary representations.

that these magnified differences enhance the anomaly-detection process.

B. Data Representation as a Binary Form of the WCs

The location and magnitude of peaks play a very important role in classifying radioactive materials because peaks are associated with either the isotopes that generate the data or the measure of radiation counts. However, peaks may be of small or large magnitude. The idea of distinguishing between small or large peaks is modeled into our approach by classifying all data points according to a set threshold. This is an extension of the popular peak-over-threshold (POT) approach. The POT approach is an extreme value theory technique that has widely been used in water resources research for peak selection (e.g., [18] and [19]). The technique involves identifying peaks that exceed a user-defined threshold. However, in our approach, instead of classifying only some of the peaks in the profile because the definition of a peak is domain dependent, we propose the classification of all the data points based on a defined threshold. This modified approach is called the local-peak-over-threshold (LPOT) if we think of each data point as a form of peak with respect to the data point in its immediate downstream and/or upstream. This is appropriate for radiation data since the entire data profile, and not only some of the peaks, is important. One challenge in using this approach is defining the threshold. Our approach in this case is to set the threshold as a measure of the global characteristic of the data. For this application, like some other applications, the mean is a good global measure for the data because there is no concern of possible outliers or influential data points and because the total count is data dependent. Hence, WCs greater than the coefficients' mean are classified as major coefficients, and other coefficients are classified as minor coefficients. To combine the classes of peaks, we adopt the idea of a binary representation.

A binary representation is a transformation process whereby a WC greater than the mean coefficient is represented by 1 and represented by 0 if otherwise. This is a form of the hard thresholding (hard limiting) approach, which has been used in the statistics community [20] and the data mining community [21], [22]. However, to our knowledge, the binary representation has not been used for representing WCs. Some of the advantages of using a binary representation have been summarized (for example, see [20]). These include a significant compression ratio while retaining a considerable amount of information in the original series and reduction in the computational cost, particularly for large time series. The advantage of a binary representation can also be seen in Fig. 7, in which the differences in WCs become distinct in the binary representation of the WC. These distinct differences in the binary domain definitely enhance the detection process, as will be shown in Sections IV-A and B.

C. Anomaly-Detection Method by a Distance-Based Metric

The proposed method is inspired by two studies in the literature [23], [24]. Our method assumes that all similar data sets have about the same mean distance from each other (this is similar to the assumption used in the agglomerative hierarchical clustering technique); therefore, they are within close proximity to one another. One key difference between our method and the previous studies, such as [23] and [24], is that our approach does not assume *fixed* proximity; rather, it bases the measure of proximity on the given data. For the proposed method, we assume that three or more data sets (for example, radiation data) are given. To use this method in a supervised manner, some data sets of normal radiation signals are given, along with the data of a new truck; the method will determine if the data of the new truck are examples of a normal instance or not. This is the most probable use of this method in transportation security applications. In an unsupervised manner, all data sets of interest are used, and the method will identify the anomalies among the set. The steps involved in the method are summarized in the list that follows.

1) *Determine the Euclidean distance between a pair of data.* The Euclidean distance between binary points taken by a pair from a list of k sets of data is

$$d_{ij} = \sqrt{(x_i - x_j)^T(x_i - x_j)} \tag{1}$$

where T is the matrix transpose, and $i, j = 1, 2, \ldots, k$.

2) *Calculate the local proximity range r_i for each data point.* This is the mean of the Euclidean distance of each data

point to other given sets of data, which is defined as

$$r_i = \frac{1}{k-1} \sum_{j=1,\, j \neq i}^{k-1} d_{ij}, \quad i, j = 1, 2, \ldots, k. \tag{2}$$

3) *Calculate the GPR*. $\text{GPR} = \text{MLP} + (Z \times \text{SLP})$, where MLP is the mean of all local proximity ranges, SLP is the standard deviation of all local proximity ranges, and Z is the *proximity factor*. Both MLP and SLP are defined as

$$\text{MLP} = \frac{1}{k^*} \sum_{i=1}^{k^*} r_i$$
$$\text{SLP} = \sqrt{\frac{1}{k^*} \sum_{i=1}^{k^*} \left[r_i - (\text{MLP}) \right]^2} \tag{3}$$

where k^* is the number of baseline data. For the unsupervised approach, $k^* = k$; for the supervised approach, $k^* < k$.

4) *Identify the anomalous data among k given sets of data.* We define the data as anomalous if their local proximity range exceeds the GPR, as follows:

$$\begin{aligned} D^*_{\text{anomaly}} &= \{ r_i \in k | r_i > \text{GPR} \} \\ D^*_{\text{normal}} &= \{ r_i \in k | r_i \leq \text{GPR} \} \end{aligned} \tag{4}$$

where D^*_{normal} are examples of usual data, and D^*_{anomaly} are examples of anomalous data.

This approach is particularly attractive because it requires only one intuitive parameter, i.e. Z, unlike most general-purpose anomaly-detection approaches that require many parameters and, of course, parameter optimization. Like other threshold-based approaches, the value of Z must cautiously be set. Based on our experience with the data sets discussed in this paper and other data sets not discussed here, setting Z to 0.5 has encouraging results and achieves zero missed detection. Therefore, we suggest that Z be set to 0.5. To justify this decision, we set Z to 0.5 and 1.0 for all analyses in this paper and compare the obtained results.

IV. Application to Transportation Security

In this section, we present an application of our proposed approach to transportation security using the simulated data described in Section II.

A. Detection of Hidden Isotopes in Radiation Data

To detect the presence of hidden isotopes in radiation data, we use the data sets in Scenario I, as shown in Fig. 3. For this case study, we set up the scenario such that three of the five data sets (radiation signals) are examples of normal data for trucks transporting kitty litter, and we compare the signals from two other trucks with these known cases to determine if they are also examples of this NORM or not. The assumption is that the cases with the SNM sources, as well as the type of SNM, are not known *a priori* as the case in the real world.

TABLE II
Results of the Hidden Isotopes Detectability Power Simulation

Metrics	Z = 0.5	Z = 1.0
Detection Rate	2:2 (cases 1& 4 correctly detected)	1:2 (case 4 correctly detected)
False Positive Rate	0:3 (no false alarm)	0:3 (no false alarm)
Missed Detection Rate	0 (no missed detection)	1:2 (case 1 missed)

A step-by-step implementation of the anomaly detection method, as presented in Section III-C, using these five cases is described in the list that follows. Each of the five cases is first transformed into WCs and then into binary representations.

1) The Euclidean distance between the five binary representations of the five cases in Scenario I is given as a lower triangular matrix, i.e.,

$$d_{ij} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1.0000 & 0 & 0 & 0 & 0 \\ 1.0000 & 0 & 0 & 0 & 0 \\ 1.0000 & 1.4142 & 1.4142 & 0 & 0 \\ 1.0000 & 0 & 0 & 1.4142 & 0 \end{bmatrix}.$$

2) The local proximity range for each data point $(r_i,\ i = 1, 2, \ldots, 5)$ is

$$r_i = \{1.0000, 0.6036, 0.6036, 1.3107, 0.6036\}.$$

3) The mean of the local proximity range, i.e., MLP, and the variability of the local proximity range, i.e., SLP, for the five data are given as follows:

$$\text{MLP} = 0.8243, \qquad \text{SLP} = 0.3216.$$

4) Setting $Z = 0.5$, $\text{GPR} = 0.9850$; therefore, $D^*_{\text{anomaly}} = \{1, 4\}$ and $D^*_{\text{normal}} = \{2, 3, 5\}$. For $Z = 1.0$, $\text{GPR} = 1.1458$; then, $D^*_{\text{anomaly}} = \{4\}$, and $D^*_{\text{normal}} = \{1, 2, 3, 5\}$.

These results are also summarized in Table II. This example is implemented in an unsupervised manner since the label of the cases is considered not known *a priori*. Assuming that we knew that Cases 2, 3, and 5 are examples of a normal occurrence, we can implement the approach in a supervised manner by repeating Steps 3 and 4.

3) The MLP and SLP for the three examples of a normal occurrence are

$$\text{MLP} = 0.6036, \qquad \text{SLP} = 0.$$

4) Setting $Z = 0.5$ and 1.0, $\text{GPR} = 0.6036$; therefore, $D^*_{\text{anomaly}} = \{1, 4\}$.

The results for the supervised implementation also correctly identify Cases 1 and 4 as anomalies. The supervised implementation is not discussed any further.

To evaluate the obtained results for the unsupervised implementation, we used two major performance indicators for the anomaly detection process in the literature [24]: 1) the detection rate and 2) the false positive rate. The detection rate is the ratio of the number of detected anomalies to the total number of

TABLE III
RESULTS OF THE MEAN- AND TOTAL COUNT-BASED METHODS

Method and Metrics	**Mean-based**	**Total count-based**	**Our approach without the WT stage**
Detection Rate	1:2 (case 1 correctly detected)	1:2 (case 1 correctly detected)	1:2 (case 1 correctly detected)
False Positive Rate	1:3 (case 5 wrongly detected)	1:3 (case 5 wrongly detected)	1:3 (case 5 wrongly detected)
Missed Detection Rate	1:2 (case 4 missed)	1:2 (case 4 missed)	1:2 (case 4 missed)

anomalies present in the data sets, whereas the false positive rate is the ratio of the total number of normal instances that were incorrectly identified as anomalies to the total number of normal instances present in the data sets. Another metric that is used is the missed detection rate, which is the ratio of missed anomalies to the number of known anomalies in the data sets.

In Table II, we see that using $Z = 0.5$, our approach correctly detects the two anomalous trucks. However, using $Z = 1.0$, only one of the two anomalous trucks (i.e., Case 4) is detected, thereby resulting in one missed detection. It is interesting to note that the obvious case (Case 1) in the original domain is not detected when Z is set to 1.0. The tradeoff in setting Z to 1.0 or greater is that there may be some missed detections. Interestingly, the number of false alarms is zero; this is a very encouraging result for this application since one of the objectives of this approach is to reduce the number of false alarms. The results further show that spectroscopy data are useful for detecting when NORM and SNM sources are placed together in a truck.

We compare our results with the mean- and total count-based methods that have been used in the literature and with using only the second and third stages of our proposed approach (i.e., *without* the WT stage). The results are shown in Table III. These results show that the three methods are only able to detect Case 1, which is obvious from the plots in Fig. 3; hence, there is one missed detection, which could be a security risk.

These other methods wrongly detected Case 5 as an anomaly when it is actually an example of a normal instance, thereby resulting in a false alarm. The results shown in Table III validate *a posteriori* our choice to transform the data into the wavelet domain before performing anomaly detection. Indeed, if we apply only the second and third stages of our proposed approach (i.e., without the WT stage), we obtain the same performance as in the mean- and total count-based approaches, namely, we detect only one of the two anomalies (the obvious case), which results in one missed detection. On the other hand, the complete approach (i.e., with the WT stage) eliminates the missed detection.

To illustrate why our approach performs better with the WT stage, we show in Figs. 8 and 9 the binary plots of these five cases with and without WT, respectively. In Fig. 8, we can visually tell why Cases 1 and 4 are correctly detected: They have extra peaks at data point 6; in addition, Case 4 has one more peak at data point 65 compared with the other plots. These extra peaks are more distinct using the WT stage, thus enhancing the detection process. In Fig. 9, on the other hand, only Case 1 is correctly detected because it has a lesser number of peaks around data point 20 compared with the other plots.

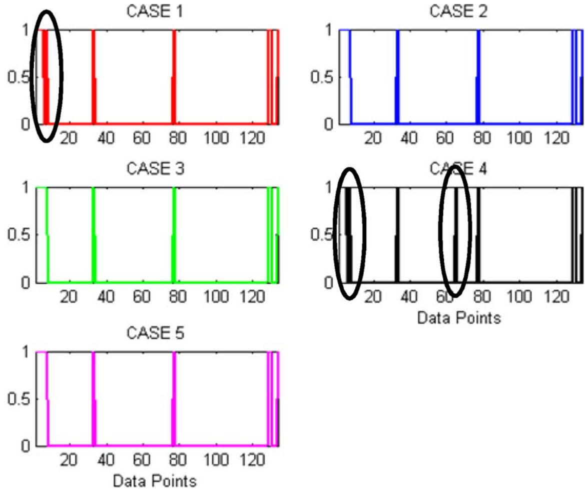

Fig. 8. Binary representation of the five cases in Scenario I using our approach.

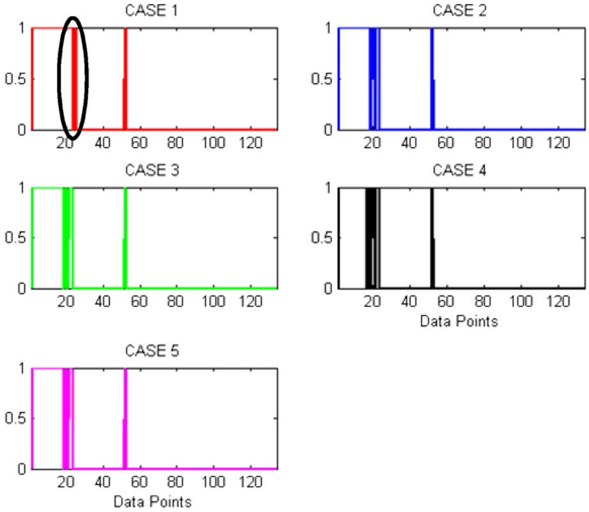

Fig. 9. Binary representation of the five cases in Scenario I using our approach without the WT stage.

This application is limited in that only one type of NORM was present in the truck; therefore, more experiments are needed to further test the capabilities of the approach in detecting hidden isotopes within other types of NORMs or TENORMs and mixture of NORMs or TENORMs. Even then, this application does offer the intriguing possibility of using our approach to detect the presence of illicit isotopes within NORMs or TENORMs.

B. Detection of the Shielded SNM and Low Percentage of the NORM

To detect the presence of a physical shielding agent with the NORM or the use of different percentages of the NORM, we present the results of the other five scenarios listed in Table I. The results (see Table IV) provide some interesting insights. The use of detection and false positive rates are not intuitive for

TABLE IV
RESULTS OF THE MINIMUM NORM DETECTABILITY POWER SIMULATION

	Gross Count Data			
	$Z = 0.5$		$Z = 1.0$	
Scenario	# of Detected Anomaly	Case(s)	# of Detected Anomaly	Case(s)
II	0	-	0	-
III	1	7	1	7
IV	0	-	0	-
V	3	1, 3, 6	3	1, 3, 6
VI	3	2, 3, 7	3	2, 3, 7
	Spectroscopy Data			
	$Z = 0.5$		$Z = 1.0$	
Scenario	# of Detected Anomaly	Case(s)	# of Detected Anomaly	Case(s)
II	2	2, 7	2	2, 7
III	1	2	1	2
IV	1	2	1	2
V	3	2, 4, 6	3	2, 4, 6
VI	0	-	0	-

these scenarios; therefore, we base our discussion on what are identified as anomalies.

Using the gross count data, our approach is not able to detect any difference between the cases that have different percentages of the NORM whether the NORM has a shielding agent embedded in it or not. In Scenarios II and IV, where all cases are with or without a metal box, no anomaly is detected. However, our approach is able to detect Case 7 in Scenario III. For Scenarios V and VI, our approach correctly detects all the cases with a box (Cases 1, 3, and 6 in Scenario V and Cases 2, 3, and 7 in Scenario VI). These results indicate that using our approach, gross data may be more useful for detecting the presence of a shielding agent such as a metal box rather than differences in the percentage of the NORM.

On the other hand, using the spectroscopy data, our approach consistently detects cases with 20% NORM or less as anomalies; this indicates that cases with at least 30% NORM are not different from cases with 100% NORM. However, none of the cases with a box is detected as an anomaly. For Scenario V, our approach detects the cases with at most 20% NORM. Of special interest is Case 6, which is detected as an anomaly using both types of data because it contains 20% NORM and a box; thus, it is an anomaly using gross data from the shielding perspective and using spectroscopy data since the percentage of the NORM is less than 30%. No case is detected as an anomaly in Scenario VI because all the cases use between 30% and 70% NORM and no box. Again, more additional experiments are needed to further justify the performance of our approach. For example, how would the approach perform in cases that contain both an SNM source and a 10% or a 20% NORM source?

We applied our approach without the WT stage (as described in Section IV-A) to these five scenarios, but the obtained results are not consistent, thereby making interpretations very difficult. This further supported our decision to transform the original data into WCs for more accurate detections and better interpretations. Interestingly, the use of $Z = 0.5$ and 1.0 give the same results, which further confirms that we can set $Z = 0.5$ in all cases without loss of generalization.

V. Implications for Transportation Security

Some of the ways adversaries may use commercial trucks to transport illicit radioactive materials include transporting them with some common NORMs or TENORMs in commerce, transporting them shielded with a thick metal box, and using these two ways together. Any of these three ways is a challenge for the existing detection process. In this paper, we have presented an anomaly detection approach for detecting trucks transporting illicit radioactive materials using each of these three possibilities.

The approach uses a simple and intuitive user-defined threshold that is suitable for cases with no prior knowledge about the materials in the trucks. The extensive experimentation using 75 simulated cases grouped into 11 scenarios indicates the potentials of the approach in detecting anomalies under various conditions. The different conditions investigated are

1) physical shielding of the SNM with a metal object;
2) using a smaller percentage of the SNM within legitimate cargo;
3) shielding of the SNM with legitimate commercial cargo;
4) shielding of the SNM with legitimate NORMs of different compositions.

The results of the various detection processes indicate that the use of gross count data alone (as it is currently being done) may not be sufficient for detecting trucks that are sources of security hazards on the highways. This supports previous conclusions in the literature that energy (spectroscopy) data are also needed for effective discrimination between SNMs and NORMs in commercial trucks. In general, concomitantly using both types of radiation data could provide more insights than separately using each of them; our approach can be used for both data types. Finally, in all considered cases, we see that our approach performs better than the mean- and total count-based methods; furthermore, the transformation of the original data to WCs significantly improves the performance of our approach.

From the perspective of smuggling illicit radioactive materials, the results show that our approach has the potential of minimizing the numerous problems associated with the security risk of illicit trucks on the highways. Specifically, our approach has the potential of reducing the number of false alarms and eliminating missed detections, which could decrease the additional costs of operating weigh stations due to secondary and tertiary inspections of trucks and prevent the transportation of illicit radioactive materials to secure and sensitive locations around the country. This would further cut down on the truck processing time at weigh stations and reduce the impacts of truck inspection on supply chain networks. This approach is suitable for implementation at border crossings, ports, and weigh stations in the country. Our approach requires the identification of some base cases to be used for comparing new trucks at weigh stations or similar facilities; the base cases can be grouped by the vehicle type and the NORM type, among others.

Since our analyses in this paper are based on simulated data and one type of NORM, there is a need for testing with real-world data, mixture of NORMs, and additional SNM sources that behave like common NORMs or TENORMs in commerce. Even then, these results do offer the intriguing possibility of using our approach to detect shielded illicit isotopes. In addition, the results provided insights on how to design the real-world data collection process: for example, the need to collect two types of radiation data rather than one type of radiation data and using different percentages of the NORM. Furthermore, the use of radiation data with other sensors and heterogeneous data should be considered to enhance the reliability of the detection process, particularly in the case of overshielding and using less than 30% NORM. For example, can an SNM source be shielded so much that it would not be detected using both types of radiation data? If yes, could the use of radiation data, along with the truck weight profile, provide more insights about the security risk of such trucks? In addition, is information about the truck driver(s) or the haulage company useful for the detection process? These questions will be addressed in future research.

Acknowledgment

The authors would like to thank R. Walker, D. Hill, and the SensorNet Program managed by the Computational Sciences and Engineering Division, Oak Ridge National Laboratory (ORNL), for providing the weigh station inspection flow diagram used in Fig. 1; the scientists from ORNL, namely, C. Liu, I. Gross, R. Walker, C. Smith, B. Gorman, and F. DeNap, for informal discussions; D. Getman and an anonymous reviewer for their comments and suggestions within the ORNL's internal review system; and two external anonymous reviewers for their constructive comments, which improved the quality of this paper. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

[1] H. Chen, F.-Y. Wang, and D. Zeng, "Intelligence and security informatics for homeland security: Information, communication, and transportation," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 4, pp. 329–341, Dec. 2004.

[2] R. T. Kouzes, J. H. Ely, B. D. Geelhood, R. R. Hansen, E. A. Lepel, J. E. Schweppe, E. R. Siciliano, D. J. Strom, and R. A. Warner, "Naturally occurring radioactive materials and medical isotopes at border crossings," in *IEEE Nucl. Sci. Symp. Conf. Rec.*, 2003, vol. 2, pp. 1448–1452.

[3] J. H. Ely, R. T. Kouzes, B. D. Geelhood, J. E. Schweppe, and R. A. Warner, "Discrimination of naturally occurring radioactive material in plastic scintillator materials," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 4, pp. 1672–1676, Aug. 2004.

[4] S. M. Brennan, A. M. Mielke, D. C. Torney, and A. B. Maccabe, "Radiation detection with distributed sensor networks," *Comput.*, vol. 37, no. 8, pp. 57–59, Aug. 2004.

[5] S. M. Brennan, A. M. Mielke, and D. C. Torney, "Radioactive source detection by sensor networks," *IEEE Trans. Nucl. Sci.*, vol. 52, no. 3, pp. 813–819, Jun. 2005.

[6] B. D. Geelhood, J. H. Ely, R. R. Hansen, R. T. Kouzes, J. E. Schweppe, and R. A. Warner, "Overview of portal monitoring at border crossings," in *IEEE Nucl. Sci. Symp. Conf. Rec.*, 2003, vol. 2, pp. 513–517.

[7] T. E. Valentine, "Overview of nuclear detection needs for homeland security," in *Proc. Conf. Amer. Nucl. Soc., PHYSOR*, Vancouver, BC, Canada, Sep. 10–14, 2006.

[8] MCNP, *A General Monte Carlo N-Particle Transport Code, Version 5, Volume I: Overview and Theory*, Apr. 2003. LA-UR-03-1987.

[9] R. N. Bracewell, *The Fourier Transform and Its Applications*, 3rd ed. Boston, MA: McGraw-Hill, 2000.

[10] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, S. H. Shih, Q. Zheng, C. C. Tung, and H. H. Liu, "The empirical mode decomposition method and the Hilbert spectrum for non-stationary time series analysis," *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, vol. 454, pp. 903–995, 1998.

[11] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.

[12] C. K. Chui, *An Introduction to Wavelets*. San Diego, CA: Academic, 1992.

[13] V. Alarcon-Aquino and J. A. Barria, "Anomaly detection in communication networks using wavelets," *Proc. Inst. Elect. Eng.—Commun.*, vol. 148, no. 6, pp. 355–362, Dec. 2001.

[14] K.-M. Lau and H. Weng, "Climate signal detection using wavelet transform: How to make time series sing," *Bull. Amer. Meteorol. Soc.*, vol. 76, no. 12, pp. 2391–2402, Dec. 1995.

[15] L. Li and G. Lee, "DDoS attack detection and wavelets," *Telecommun. Syst.*, vol. 28, no. 3/4, pp. 435–451, Mar. 2005.

[16] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[17] B. Vidakovic and P. Muller, *Wavelets for Kids: Tutorial Introduction*. Durham, NC: Duke Univ., Inst. Statist. Decision Sci., 1994. Discussion Paper 94-13.

[18] P. Claps and F. Laio, "Can continuous streamflow data support flood frequency analysis? An alternative to the partial duration series approach," *Water Resour. Res.*, vol. 39, no. 8, pp. 1216–1226, Aug. 2003.

[19] M. Lang, T. B. M. J. Ouarda, and B. Bobée, "Towards operational guidelines for over-threshold modeling," *J. Hydrol.*, vol. 225, no. 3/4, pp. 103–117, Dec. 1999.

[20] B. Kedem and E. Slud, "On goodness of fit of time series models: An application of higher order crossings," *Biometrika*, vol. 68, no. 2, pp. 551–556, Aug. 1981.

[21] A. Bagnall, C. Ratanamahatana, E. Keogh, S. Lonardi, and G. Janacek, "A bit level representation for time series data mining with shape based similarity," *Data Mining Knowl. Discovery*, vol. 13, no. 1, pp. 11–40, Jul. 2006.

[22] C. A. Ratanamahatana and E. Keogh, "Three myths about dynamic time warping," in *Proc. SIAM Int. Conf. Data Mining*, Newport Beach, CA, 2005, pp. 506–510.

[23] P. K. Chan, M. V. Mahoney, and M. H. Arshad, "Learning rules and clusters for anomaly detection in network traffic," in *Managing Cyber Threats: Issues, Approaches and Challenges*, V. Kumar, J. Srivastava, and A. Lazarevic, Eds. New York: Springer-Verlag, 2005, pp. 81–99.

[24] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data," in *Applications of Data Mining in Computer Security*, D. Barbara and S. Jajodia, Eds. New York: Kluwer, 2002.

Olufemi A. Omitaomu (S'05–M'07) received the B.Sc.(Hons.) degree in mechanical engineering from Lagos State University, Ojo, Nigeria, in 1995, the M.Sc. degree in mechanical engineering from the University of Lagos, Lagos, Nigeria, in 1999, and the Ph.D. degree in industrial and information engineering from the University of Tennessee, Knoxville, in 2006.

He is currently with the Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN. From 1995 to 2001, he worked in the oil and gas industry as a Project Engineer. He is the coauthor of *Computational Economic Analysis for Engineering and Industry* (CRC, 2007) and the coeditor of *Knowledge Discovery from Sensor Data* (CRC, 2009). His research interests include data mining and knowledge discovery from sensors and streams, electric power-grid modeling and analysis, critical infrastructure interdependences, risk analysis in space and time, optimization, and applied artificial intelligence.

Dr. Omitaomu is a member of the Institute for Operations Research and Management Sciences.

Auroop R. Ganguly (M'99) received the B.Tech.(Hons.) degree in civil engineering from the Indian Institute of Technology, Kharagpur, India, in 1993, the M.S. degree in civil engineering from the University of Toledo, Toledo, OH, in 1997, and the Ph.D. degree in civil and environmental engineering, with a concentration in hydrology, from the Massachusetts Institute of Technology, Cambridge, in 2002.

He is currently a Research Scientist within the Computational Sciences and Engineering Division, Oak Ridge National Laboratory (ORNL), Oak Ridge, TN. He is also an Adjunct Professor with the University of Tennessee, Knoxville. Prior to joining ORNL, he had more than five years of experience in the software industry, specifically with Oracle Corporation and a best-of-breed company subsequently acquired by Oracle, and about a year in academia, specifically with the University of South Florida, Tampa. He has published peer-reviewed papers in hydrological, meteorological, climate, and nonlinear dynamics journals and has presented at peer-reviewed computer science conferences and workshops. He is the coeditor of *Knowledge Discovery From Sensor Data* (CRC, 2009). His research interests are climate change and extremes, hydrology and hydrometeorology, complex systems analysis, geoscience informatics, computational data sciences, and space–time knowledge discovery.

Dr. Ganguly is a member of the American Geophysical Union, the American Meteorological Society, and Sigma Xi. He is currently a member of the invited reader panel of the journal *Nature*. He has been invited to panels and workshops organized by the National Science Foundation, holds an Associate Editor appointment with the *Journal of Computing in Civil Engineering* published by the American Society of Civil Engineers, and has participated in workshops organized by the U.S. Department of Energy and the U.S. Department of Homeland Security. His research has been publicized in the media and at scientific venues.

Bruce W. Patton received the B.S. degree in applied physics from Louisiana Tech University, Ruston, the M.S. degree in nuclear engineering from the Georgia Institute of Technology, Atlanta, and the Ph.D. degree in nuclear engineering from the University of Michigan, Ann Arbor.

He is currently with the Nuclear Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN. He has served as a Nuclear Propulsion Officer with the U.S. Navy and has worked for commercial industry and the National Aeronautics and Space Administration. His research interests include radiation transport computational methods and simulations, molten salt reactors, and applications of genetic algorithms to nuclear engineering.

Vladimir A. Protopopescu received the Ph.D. degree in mathematical physics from the Institute for Atomic Physics, Bucharest, Romania, in 1976.

In 1985, he joined the Oak Ridge National Laboratory, Oak Ridge, TN, where he is currently the Chief Scientist with the Computational Sciences and Engineering Division. From 1968 to 1984, he successively worked with the Institute for Atomic Physics, Chalmers University of Technology, Göteborg, Sweden, Yale University, New Haven, CT, and Boston University, Boston, MA. He serves as an Associate Editor of the journal *Transport Theory and Statistical Physics* and of two book series *Modeling and Simulation in Science, Engineering, and Technology* (Birkhauser) and *Mathematical Modeling* (CRC). He is the holder of several U.S. patents. His research interests include mathematical modeling, analysis and optimal control of partial differential equations, dynamical systems, inverse problems, global optimization, and modern application of the control theory to quantum systems.

Dr. Protopopescu is a member of the American Mathematical Society, the Society for Industrial and Applied Mathematics, the American Physical Society, and the International Association of Mathematical Physics. He was the recipient of the R&D 100 Awards for his work on global optimization and on time series analysis of electroencephalogram signals in 1998 and 2005, respectively.