Automated Detection of Anomalous Shipping Manifests to Identify Illicit Trade

Antonio Sanfilippo and Satish Chikkagoudar
Computational & Statistical Division, National Security Directorate
Pacific Northwest National Laboratory
Richland, WA, USA
{antonio.sanfilippo, satish.chikkagoudar}@pnnl.gov

Abstract—We describe an approach to analyzing anomalies in trade data based on the identification of cluster outliers. The approach uses unsupervised machine learning methods to discover semantically coherent clusters of shipping records in large collections of trade data. Trade data with cluster annotations are then used as input to a supervised machine learning algorithm to train and evaluate a classification model capable of identifying members of each cluster. The evaluation of this classification model provides an assessment of cluster coherence. Outliers are identified for each cluster by measuring the Euclidean distance from each member of the cluster to the cluster centroid, and then selecting a percentile threshold to identify shipping records with extreme distances from the cluster centroid. We describe a specific application of this approach to a dataset of 2.36M records for containerized shipments, with specific reference to the detection of anomalies potentially related to nuclear smuggling. Results show that this approach succeeds in finding semantically coherent clusters of shipping records, and identifying outliers that may help facilitate the detection of illicit trade.

Keywords—trade data; clustering; classification; visual analytics; illicit trafficking; detection of radiological threat materials; nuclear smuggling.

I. INTRODUCTION

About 90% of world trade is carried by the international shipping industry, with more than 50,000 merchant ships trading internationally, registered in over 150 nations, and operated by over a million seafarers across the world [1]. The opportunities for illicit trade are numerous and often difficult to prevent or seize. Consider, for example, the illicit trafficking of nuclear materials. The use of radiation detection instrumentation [2,3] increases the ability to detect nuclear materials shipped illicitly. However, the ubiquity of naturally occurring radioactive materials that are known to trigger false alarms [4] constitutes a major challenge for the successful application of radiation detection techniques [5, 6]. Manual inspection of all shipments which trigger an alarm would implement a strong interdiction policy, but it is not feasible as it would disrupt the regular flow of trade with intolerable economic consequences.

Recent studies [6,7] have shown that the analysis of shipping manifests for containerized commodities can help improve the accuracy of physical interrogation techniques to

detect radiological threat materials in cargo at US borders. In this paper we extend the reach of these approaches by using machine learning methods to infer and evaluate clustering and classification models from historical shipping manifest data that help identify traits of new transactions that may lead to the detection of illicit nuclear trade. We also show how the insights gained through the inferred models can be brought to the user through visual analytic capabilities to support situation awareness in dynamic decision making in monitoring and warning against the movement of radiological threat materials.

We begin by providing additional background details about the problem addressed, and surveying past related work to better frame our approach. Next, we describe the data used, the modeling techniques applied to the data, and the results obtained. We conclude with a description of the analytic methods used to bring the insights borne out by the models to the user, and a brief overview of next steps.

II. BACKGROUND

The illegal shipment of radiological threat materials is a major threat to national security. The U. S. Customs and Border Protection has been deploying radiation portal monitors (RPMs) which provide a passive, non-intrusive means to screen trucks and other conveyances for the presence of nuclear and radiological materials [8]. RPMs detect traces of radiation emitted from an object passing through an RPM. Gamma radiation is detected, and in some cases complemented by neutron detection to enable sensitivity for nuclear materials [5]. Other detection techniques include active interrogation, for example by probing a container with a neutron beam to detect beta delayed gamma rays emitted by fission products [3].

A major challenge for the successful deployment of radiation detection techniques is the ubiquity of Naturally Occurring Radioactive Materials (NORM) and Technologically Enhanced Naturally Occurring Radioactive Materials (TENORM) for both passive and active detection of radioactive material [4, 5, 6]. For example, fluorine provides a main source of interference for neutron active interrogation techniques [6]. Fluorinated commodities, compounds and materials may therefore trigger false positive radioactive

detection [6]. The same holds for passive interrogation methods, for which false alarms are triggered by commodities such as fertilizer, ceramic glazed materials, aircraft parts, camera lenses, polishing compounds and abrasives, fluorescent lamp starters, welding rods, propane tanks, kitty litter, road salt, ore and rock, medical isotopes, and smoke detectors [5].

The concurrent analysis of bills of lading data can greatly improve the detection capability of physical interrogation methods by profiling commodity type by shipment to reduce false alarm. The approaches developed so far [6, 7] have primarily been descriptive by illustrating correlations between NORM and TENORM commodities and diverse parameters of shipping manifests, e.g. ports of arrival, imported commodities, cargo composition. In this paper, we show how these correlations can be brought to bear in a more effective way by engaging data mining techniques such as clustering and classification to automate the identification of anomalies. These enhancements offer an operational way to complement radiation detection techniques to detect undeclared traffic of nuclear materials.

III. MODELING TRADING DATA

A. Description of the shipping manifest data used

The data used in this study were obtained from the Port Import Export Reporting Service (PIERS). PIERS collects more than 15 million bills of lading per year relative to over 20 million shipments for the US, most of which are containerized commodities. PIERS processes these data into databases, facts, and figures, which others can then use to understand global trade. All commodity data used in this study were from calendar year 2006; see [7] for a more detailed description of these data including transformations applied to the initial resources.

The full PIERS dataset for 2006 includes several million shipping entries. For our exploratory study we selected the first quarter of 2006, which contains over 2.36 million entries. Our dataset contains information about 19 fields (Table 1) out of approximately 75 potential fields that appear in the full PIERS record [7].

B. Clustering

We used the expectation maximization (EM) clustering algorithm [9] to group the 2.36 million shipping records described in section III.A into 25 clusters. The EM algorithm assumes that the samples in a dataset are drawn from several different probability distributions. It generates the probabilities that measure the extent to which a sample belongs to the various clusters. Cluster membership is decided by probabilities in contrast to K-Means algorithm which utilizes Euclidean distance [10] for that purpose. We chose the EM algorithm because we found it more accurate than *k*-means, and still capable of dealing with the substantial number of records (2.36 million) as compared to more computationally intensive algorithms (e.g. hierarchical clustering). Also, the EM algorithm is designed to capture correlations and dependences of attributes. As such, it

provides an ideal analysis of shipping records for the characterization of anomalies that may signal illicit trade in terms of unlikely associations of shipping properties (e.g. illogical routes and buyers). Such anomalies are related to the red flag indicators issued by the Bureau of Industry and Security at the U.S. Department of Commerce [11], which provide a checklist to discover possible violations of the Export Administration Regulations [12].

SLINE Ship Line-Carrier Code VESSEL Name of the waterborne vessel VOYAGE Voyage number YRMTDY Date field HSCODE Harmonized tariff code for all products and commodities HARM DESC HSCODE description USCODE Port in the U.S. that cargo crossed USPORT USCODE description FCODE Foreign port code FPORT FCODE description ULTCODE Port of origin (U.S. imports) or destination (U.S. exports) code ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containers TEUS Twenty Foot Equivalent – Units MTONS Metric Tons			
VOYAGE Voyage number YRMTDY Date field HSCODE Harmonized tariff code for all products and commodities HARM_DESC HSCODE description USCODE Port in the U.S. that cargo crossed USPORT USCODE description FCODE Foreign port code FPORT FCODE description ULTCODE Port of origin (U.S. imports) or destination (U.S. exports) code¹ ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containers TEUS Twenty Foot Equivalent – Units	SLINE	Ship Line-Carrier Code	
YRMTDY Date field HSCODE Harmonized tariff code for all products and commodities HARM DESC HSCODE description USCODE Port in the U.S. that cargo crossed USPORT USCODE description FCODE Foreign port code FPORT FCODE description ULTCODE Port of origin (U.S. imports) or destination (U.S. exports) code ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containerized cargo flag CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	VESSEL	Name of the waterborne vessel	
HSCODE Harmonized tariff code for all products and commodities HARM DESC HSCODE description USCODE Port in the U.S. that cargo crossed USPORT USCODE description FCODE Foreign port code FPORT FCODE description ULTCODE Port of origin (U.S. imports) or destination (U.S. exports) code ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containerized cargo flag CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	VOYAGE	Voyage number	
commodities HARM_DESC HSCODE description USCODE Port in the U.S. that cargo crossed USPORT USCODE description FCODE Foreign port code FPORT FCODE description ULTCODE Port of origin (U.S. imports) or destination (U.S. exports) code ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containerized cargo flag CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	YRMTDY	Date field	
HARM DESC HSCODE description USCODE Port in the U.S. that cargo crossed USPORT USCODE description FCODE Foreign port code FPORT FCODE description ULTCODE Port of origin (U.S. imports) or destination (U.S. exports) code ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containers TEUS Twenty Foot Equivalent – Units	HSCODE	Harmonized tariff code for all products and	
USCODE Port in the U.S. that cargo crossed USPORT USCODE description FCODE Foreign port code FPORT FCODE description ULTCODE Port of origin (U.S. imports) or destination (U.S. exports) code ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containers are CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units		commodities	
USPORT USCODE description FCODE Foreign port code FPORT FCODE description ULTCODE Port of origin (U.S. imports) or destination (U.S. exports) code ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containers are CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	HARM_DESC	HSCODE description	
FCODE Foreign port code FPORT FCODE description ULTCODE Port of origin (U.S. imports) or destination (U.S. exports) code ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containers are CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	USCODE	Port in the U.S. that cargo crossed	
FPORT FCODE description ULTCODE Port of origin (U.S. imports) or destination (U.S. exports) code ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containers are CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	USPORT	USCODE description	
ULTCODE Port of origin (U.S. imports) or destination (U.S. exports) code¹ ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containerized cargo flag CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	FCODE	Foreign port code	
(U.S. exports) code ¹ ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containerized cargo flag CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	FPORT	FCODE description	
ULTPORT ULTCODE description CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containerized cargo flag CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	ULTCODE	Port of origin (U.S. imports) or destination	
CTRYCODE Country of origin code COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containerized cargo flag CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units		(U.S. exports) code ¹	
COUNTRY CTRYCODE description CONSIZE Container size CONFLAG Containerized cargo flag CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	ULTPORT	ULTCODE description	
CONSIZE Container size CONFLAG Containerized cargo flag CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	CTRYCODE	Country of origin code	
CONFLAG Containerized cargo flag CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	COUNTRY	CTRYCODE description	
CONQTY Number of Containers TEUS Twenty Foot Equivalent – Units	CONSIZE	Container size	
TEUS Twenty Foot Equivalent – Units	CONFLAG	Containerized cargo flag	
	CONQTY	Number of Containers	
MTONS Metric Tons	TEUS	Twenty Foot Equivalent – Units	
	MTONS	Metric Tons	

Table 1: PIERS filed codes in the dataset used in this study.

Cluster sizes are shown in Figure 1. With the exception of cluster 6 which contains far more records than the other clusters, cluster sizes range from 45K to 10K records. We have been experimenting with higher cluster numbers (up to 75) with results rivaling those reported in this paper. A higher number of clusters may achieve a more balanced record distribution in terms of cluster sizes.

We used information gain [13] to establish the contribution of the PIERS codes to clustering. The information gain of a given attribute X (a field code in the PIERS data) with respect to a class attribute C (one of the 25 clusters in Figure 1) is the reduction in entropy of the value for C when we know the value of X. So, by calculating the information gain of each PIERS field code with respect to the 25 clusters, we can quantify the contribution of the PIERS code to the clustering process. We used the Weka implementation of the information gain algorithm [14] to perform this calculation. After removing redundant codes (i.e. HARM_DESC, USPORT, FPORT, ULTPORT, COUNTRY, which are repeated as numerical codes HSCODE, USCODE, FCODE,ULTCODE, and CTRYCODE), we obtained the information gain weights shown in Figure 2.

530

¹ All records in the PIERS datasets used in this study are for U.S. imports.

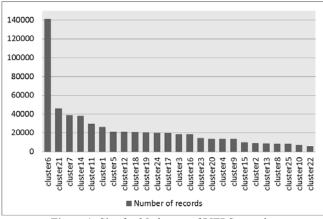


Figure 1: Size for 25 clusters of PIERS records.

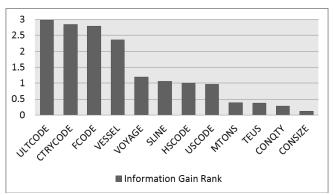


Figure 2: Information gain rankings of PIERS codes for 25 clusters.

C. Classification

To verify the internal semantic cohesion of each of the 25 clusters and the semantic distinctiveness of each cluster with respect to the others, we treated the PIERS dataset with the added cluster annotations as a training dataset for classification. Each cluster was treated as a class. We used the Weka Bayes net classifier [15] to build a model that would recognize members of the 25 cluster classes in terms of PIERS codes. We chose the Bayes net classifier because it provided superior performance as compared to other classifiers in terms of accuracy and computability. We removed the redundant codes (see section III.B), thus reducing the number of codes used for classification from 19 to 12 (see Figure 2 and related discussion). We trained on 66% of the data (1.576 million records) and tested on the remaining 34%. Evaluation results by cluster are provided in Figure 3. Overall, the results are very good, with weighted averages in precision, recall, F1measure and Receiver Operating Characteristic (ROC) area all above 90%. In addition to providing an evaluation for the 25 clusters of PIERS records, this Bayes net classifier also (and most importantly) offers an effective tool for identifying properties of new shipping manifests by association with one of the 25 clusters.

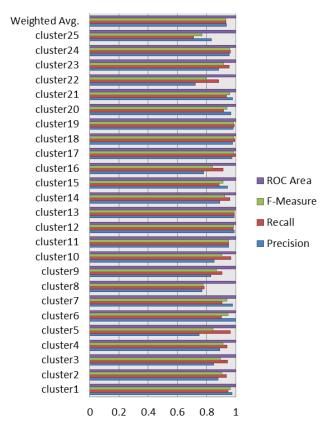


Figure 3: Evaluation results for 25 clusters of PIERS data records.

D. Idendification of Cluster Outliers

Cluster outliers can be identified by generating centroids for each cluster and then measuring the distance of each manifest record in the cluster from the cluster centroid. We used the Weka implementation of the K-Means algorithm [16] to perform this operation.

The K-means clustering algorithm uses the Euclidean distance of each data sample from the centroids to define cluster membership. Euclidean distance is a numerical quantity and categorical variables are converted to numerical quantities before the distance calculation. Euclidean distance between two data samples A and B is given by the formula $\left|\sqrt{\sum_{i=1}^{n}(A_i-B_i)^2}\right|$ [10]. This analysis is exemplified below with reference to cluster 2.

Table 2 shows the content of the centroid for cluster 2, as identified with the K-Means algorithm. The distribution of each shipping record in cluster 2 is determined by the Euclidean distance [10] of the record from the centroid. Our hypothesis is that by selecting a high Euclidean distance threshold, we can identify anomalous records which present a higher risk for illicit trafficking due to their unusual combination of cosigner, consignee, carriers, ports of departure/arrival, and other properties.

IV. ANALYSIS OF ANOMALOUS SHIPPING RECORDS FOR NORM AND TENORM COMMODITIES

We analyzed the PIERS data enriched with cluster and centroid information using Spotfire [17]. Spotfire is a software platform that allows users to analyze data using visual analytic methods. It includes the ability to develop dynamic analytic applications that run on the web through a client called TIBCO Spotfire Web Player.

CODE	VALUE	CODE	VALUE
SLINE	SLINE I	ULTCODE	41585
VESSEL	VESSEL J	ULTPORT	VALPARAISO
VOYAGE	VOYAGE K	CTRYCODE	415
YRMTDY	060215	COUNTRY	CHILE
HSCODE	243354	CONSIZE	20
HARM_DESC	GRAPES, FRESH	CONFLAG	С
USCODE	2019	CONQTY	0.158
USPORT	PHILADELPHIA	TEUS	0.240599
FCODE	40991	MTONS	5379.813
FPORT	VALPARAISO		

Table 2: Code-value pairs forming the centroid for cluster 2. SLINE, VESSEL and VOYAGE codes have been anonymized to ensure data privacy.

Using Spotfire, we formed a query to retrieve all clusters with records where the code for the harmonized tariff description of commodities (HARM_DESC) contained one of the terms for commodities which trigger false alarms with radiation portal monitors:

fertilizer, ceramic glazed materials, aircraft parts, camera lenses, polishing compounds and abrasives, fluorescent lamp starters, welding rods, propane tanks, kitty litter, road salt, ore and rock, medical isotopes, and smoke detectors.

The ensuing results, shown in Figure 4, exemplify clusters containing NORM/TENORM commodities. These contain 8800 shipping records and constitute about 0.38% of the entire dataset (2.36M records).

While shipping records for NORM/TENORM commodities form a small percentage of our entire dataset, these still constitute a large number of shipments to select for manual inspection: about 146 containers per day.² Also, consider that shipments for other commodities would also have to be inspected to check for other types of illicit trade (e.g. fire arms, drugs, humans, art and antiquities, etc.) We can drastically reduce the number of shipping records for NORM/TENORM commodities to be selected for further inspection by restricting the selection to anomalous shipping records. For example, when we analyze the largest cluster of shipping records for NORM/TENORM commodities for outliers (Figure 5), we find that only 16 out of 2540 records stand out as clear outliers. These have Euclidean distance values from the cluster centroid equal to or greater than the 99.95 percentile. Similar distributions can be observed with the other clusters, as shown is Figure 6 for the next 7 largest clusters of shipping records for NORM/TENORM commodities (Figure 4).

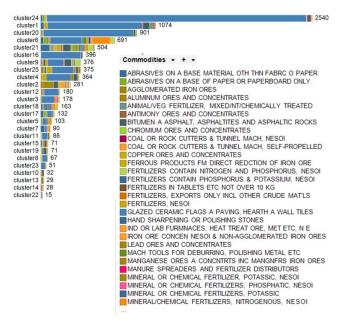


Figure 4: Results of query for NORM and TENORM commodities.

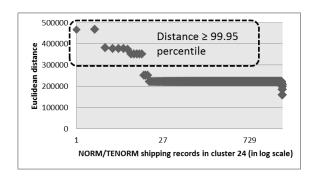


Figure 5: Analysis of outliers for the NORM/TENORM version of cluster 24.

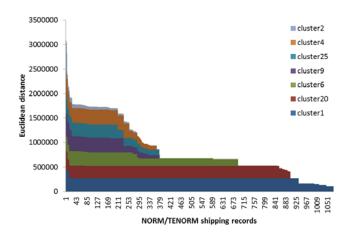


Figure 6: Analysis of outliers for NORM/TENORM clusters with more than 200 records.

² Since our dataset covers a period of three months, we are assuming that manual inspection of the 8800 shipments would be spread over 60 business days, with an average of 146 inspections per day.

Table 3 shows the list of outliers for the NORM/TENORM version of cluster 24. Not all outliers can be regarded as true cluster anomalies. For example, the first two entries in Table 3 describe commodities that occur with extreme low frequency over the entire dataset, not just cluster 24 – i.e. only 19 of the 2.36 million records are shipments of "precious metal ores and concentrates". Other entries in Table 3 describe commodities that occur with greater frequency in other clusters, which may even be smaller (e.g. cluster 21, see Figure 4), as shown for the commodity "parts of mach for sorting etc. earth stone ores etc." in Figure 7. In these cases, the relative lower frequency of the commodity within the cluster indicates that the combination of the commodity with the other field codes presents a true cluster anomaly, and as such it may deserve to be considered for further inspection.

COMMODITIES	US PORT	FOREIGN PORT	COUNTRY
PRECIOUS METAL ORES AND CONCENTRATES	NEW YORK	BUENOS AIRES	ARGENT
PRECIOUS METAL ORES AND CONCENTRATES	NEW YORK	BUENOS AIRES	ARGENT
PARTS OF MACH FOR SORTING ETC EARTH STONE ORES ETC	HOUSTON	SANTOS	BRAZIL
PARTS OF MACH FOR SORTING ETC EARTH STONE ORES ETC	NEW YORK	BUENOS AIRES	ARGENT
PARTS OF MACH FOR SORTING ETC EARTH STONE ORES ETC	NEW YORK	SANTOS	BRAZIL
PARTS OF MACH FOR SORTING ETC EARTH STONE ORES ETC	NEW YORK	SANTOS	BRAZIL
MANURE SPREADERS AND FERTILIZER DISTRIBUTORS	GALVESTON	PUNTA MANZANI	BELGIUM
ROCK DRILL OR EARTH BORE TOOLS, WORK PT NESOI, PTS	NEW YORK	FREEPORT	ECUADOR
ROCK DRILL OR EARTH BORE TOOLS, WORK PT NESOI, PTS	HOUSTON	IQUITOS	PERU
ROCK DRILL OR EARTH BORE TOOLS, WORK PT NESOI, PTS	HOUSTON	IQUITOS	PERU
ROCK DRILL OR EARTH BORE TOOLS, WORK PT NESOI, PTS	NEW YORK	SANTOS	BRAZIL
ROCK DRILL OR EARTH BORE TOOLS, WORK PT NESOI, PTS	NEW YORK	SANTOS	BRAZIL
FERROUS PRODUCTS FM DIRECT REDCTION OF IRON ORE	PHILADELPHIA	SEPETIBA BAY	BRAZIL
FERROUS PRODUCTS FM DIRECT REDCTION OF IRON ORE	PHILADELPHIA	SEPETIBA BAY	BRAZIL
FERROUS PRODUCTS FM DIRECT REDCTION OF IRON ORE	PHILADELPHIA	SEPETIBA BAY	BRAZIL

Table 3: List of outliers in the NORM/TENORM version of cluster 24 (distance ≥ 99.95 percentile).

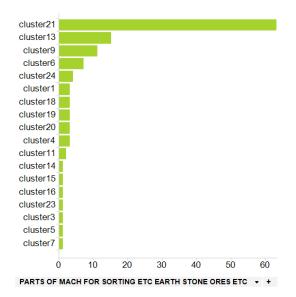


Figure 7: Frequency of commodity "parts of mach for sorting etc. earth stone ores etc." across all clusters.

V. ANALYSIS OF ANOMALOUS SHIPPING RECORDS FOR ALL COMMODITIES

The same approach described in the previous section can be used to provide a generalized characterization of illicit trade by removing the NORM/TENORM filter on shipping records. For example, the full version of cluster 2 - i.e. with entries for all commodities within the cluster - has 36K records (as opposed to the 281 records in the NORM/TENORM version of the cluster, see Figure 4). As shown in Figure 8, the Euclidean distances of each member of the cluster from the cluster's centroid are more evenly distributed as compared to the version of the cluster restricted to NORM/TENORM commodities (Figure 6). Still, if we select the same high percentile threshold (99.95) as for the NORM/TENORM of the cluster, only a handful of shipping records is selected for further inspection. As shown in table 4, most of the top outliers for cluster 2 are related to shipments of toy commodities. Interestingly, toy-making has been used as a cover to purchase dual use commodities, as in a recent case in which a Chinese merchant posing as a toy-maker allegedly sought U.S. technology for Iran [18]. So, this more generalized characterization of outliers in clusters of shipping records may also help in the detection of illicit nuclear trade.

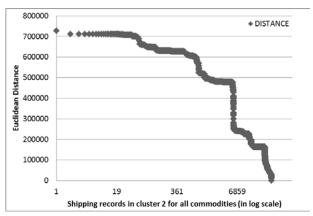


Figure 8: Distribution of the 36K records in cluster 2 in terms of Euclidean distance from the centroid (Table 2).

COMMODITY	US PORT	FOREIGN PORT
ORIGINAL SCULPTURES AND STATUARY, IN ANY MATERIAL	NEW WESTMINST	AUCKLAND (NZ)
ART FOR SPORTS ETC.NESOIF SWIM POOLSF PTS & ACCES	BALTIMORE	HONG KONG
ART FOR SPORTS ETC.NESOIF SWIM POOLSF PTS & ACCES	BALTIMORE	HONG KONG
TOYS AND PARTS AND ACCESSORIES, NESOI	BALTIMORE	HONG KONG
TOYS AND PARTS AND ACCESSORIES, NESOI	LOS ANGELES	YANTIAN (CN)
ART FOR SPORTS ETC.NESOIF SWIM POOLSF PTS & ACCES	BALTIMORE	HONG KONG
TOYS AND PARTS AND ACCESSORIES, NESOI	VANCOUVER BC	HONG KONG
TOYS AND PARTS AND ACCESSORIES, NESOI	BALTIMORE	HONG KONG
TOYS AND PARTS AND ACCESSORIES, NESOI	TACOMA	YANTIAN (CN)
ART FOR SPORTS ETC.NESOIF SWIM POOLSF PTS & ACCES	VANCOUVER BC	HONG KONG
BALLS, OTHER THAN GOLF, TENNIS AND INFLATABLE BALL	TACOMA	HONG KONG
GAME MACH EX COIN; BD GAME; MAH-JONG; DMNOES; DICE	BALTIMORE	HONG KONG
FISHING REELS, AND PARTS AND ACCESSORIES	TACOMA	HONG KONG
TOYS AND PARTS AND ACCESSORIES, NESOI	TACOMA	HONG KONG
BALLS, OTHER THAN GOLF, TENNIS AND INFLATABLE BALL	TACOMA	HONG KONG
TOYS AND PARTS AND ACCESSORIES, NESOI	BALTIMORE	YANTIAN (CN)
TOYS AND PARTS AND ACCESSORIES, NESOI	VANCOUVER BC	HONG KONG
TOYS AND PARTS AND ACCESSORIES, NESOI	VANCOUVER BC	HONG KONG

Table 4: List of outliers (Euclidean distance ≥710630.94, 99.95 percentile) for cluster 2 – only few attributes are shown.

VI. CONCLUSION

The analysis of trade data such as shipping manifests provides insights that can sharpen the accuracy of physical

interrogation methods in identifying radiological material shipped illicitly via intermodal transportation. In this paper, we have shown how analytical methods based on the development of clustering and classification models from historical shipping manifests can be used to identify anomalous shipping records that may represent instances of illicit trade. Historical data are first enriched with cluster annotations introduced via unsupervised clustering techniques such as the EM algorithm. Classification models built on these enriched data are then used to process current shipping records to assign a cluster annotation to unseen records. Once a shipping record is assigned to a cluster, its anomaly value can be computed as the Euclidean distance of the vector for the entry from the vector for the cluster's centroid.

Because the clustering algorithm used (EM) is designed to capture correlation and dependence of attributes, shipping record anomalies emerge from the lack correlation and dependence across field codes. These anomalies are therefore capable of characterizing red flag indications such as those issued by the Bureau of Industry and Security, U.S. Department of Commerce which rely on unusual combinations of parameters for shipped goods [11].

The ability to identify anomalous shipping records makes it possible to winnow down candidates for manual inspection from several hundred or thousands to just a few tens. The ensuing framework can be further enhanced by taking into consideration additional codes in shipping manifests. Moving forward, we plan to use techniques to mine social networks from online data sources to link trading agents listed in shipping manifests to individuals and organizations with restricted or denied trading rights [19-21] to enhance further the ability to identify suspicious transactions.

ACKNOWLEDGMENT

This research was supported by the Lab Directed Research & Development project on Network Analysis and Modeling of Illicit Nuclear Trafficking at the Pacific Northwest National Laboratory. The Pacific Northwest National Laboratory is managed for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC05-76RL01830.

REFERENCES

- [1] http://www.ics-shipping.org/shippingfacts (visited on 3/11/2-13).
- [2] Kouzes, RT, JH Ely, PE Keller, RJ McConn, ER Siciliano. "Passive Neutron Detection for Interdiction of Nuclear Material at Borders," NIM-A 584, 383-400. 2008.

- [3] Hall, J.M., S. Asztalos, P. Biltoft, J. Church, M.-A. Descalle, T. Luu, D. Manatt, G. Mauger, E. Norman, D. Petersen b, J. Pruet, S. Prussin b, D. Slaughter. "The Nuclear Car Wash: Neutron interrogation of cargo containers to detect hidden SNM," Nuclear Instruments and Methods in Physics Research B 261, pp. 337–340. 2007.
- [4] Radiation Information Network. Radioactivity in Nature. http://www.physics.isu.edu/radinf/natural.htm. Idaho State University. Accessed 3-12-2013.
- [5] Kouzes, R.; Ely, J.; Evans, J.; Hensley, W.; Lepel, E.; McDonald, J.; Schweppe, J.; Siciliano, E.; Strom, D.; Woodring, M. "Naturally occurring radioactive materials in cargo at US borders," Packaging, Transport, Storage and Security of Radioactive Material, Volume 17, Number 1, pp. 11-17(7). 2006.
- [6] Descalle, Marie-Anne A., Manatt, Douglas R.; Slaughter, Dennis R. "Analysis of Manifests for Containerized Commodities Imported through US Ports," 2006 IEEE Nuclear Science Symposium. 2006.
- [7] Baciak, J E, Ely, J H, Schweppe, J E, Sandness, G A, Robinson. "Analysis of International Commodity Shipping Data and the Shipment of NORM to the United States," Pacific Northwest National Laboratory, report No. PNNL-21088. 2011.
- [8] Radiation Portal Monitors Safeguard America from Nuclear Devices and Radiological Materials. U. S. Customs and Border Protection. http://www.cbp.gov/xp/cgov/border_security/port_activities/cargo_exa m/rad portal1.xml (visited on 3/11/2-13).
- [9] Wu, Cai, Jeffrey R. Steinbauer, and Grace M. Kuo. "Em clustering analysis of diabetes patients basic diagnosis index." AMIA Annual Symposium Proceedings. Vol. 2005. American Medical Informatics Association, 2005.
- [10] http://www.statsoft.com/textbook/cluster-analysis/.
- [11] www.bis.doc.gov/complianceandenforcement/redflagindicators.htm.
- [12] www.bis.doc.gov/policiesand regulations.
- [13] Mitchell, T. M. "Machine Learning", The Mc-Graw-Hill Companies, Inc. 1997. Ch. 3.
- [14] http://weka.sourceforge.net/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html.
- [15] http://weka.sourceforge.net/doc/weka/classifiers/bayes/packagesummary.html.
- $\begin{array}{ll} \textbf{[16]} & \underline{\text{http://weka.sourceforge.net/doc.dev/weka/clusterers/SimpleKMeans.htm}} \\ \underline{1} & \\ \end{array}$
- [17] Spotfire, TIBCO Software, http://spotfire.tibco.com (visited on 3/11/2-13).
- [18] Warrick, J. 2012. Nuclear ruse: Posing as toymaker, Chinese merchant allegedly sought U.S. technology for Iran. Washington Post, August 11, 2012, http://articles.washingtonpost.com/2012-08-11/world/35490055_1 nuclear-program-civilian-energy-program-chinese-firms.
- [19] Denied Persons List, Bureau of Industry and Security, http://www.bis.doc.gov/dpl/default.shtm.
- [20] Specially Designated Nationals and Blocked Persons List, Department of the Treasury, http://www.treasury.gov/resource-center/sanctions.
- [21] List of Statutorily Debarred Parties, Department of State, http://www.pmddtc.state.gov/compliance/debar.html.