Los Alamos
NATIONAL LABORATORY
EST.1943

LA-UR-18-27721
Approved for public release; distribution is unlimited.

Title: Determining Patterns from Radiation Portal Monitor Data: Enabling Data Insight with Visual and Interactive Exploratory Data Analysis

Author(s): Duncan, James P. C

Intended for: Report

Issued: 2018-08-13

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Determining Patterns from Radiation Portal Monitor Data:

Enabling Data Insight with Visual and Interactive Exploratory Data Analysis

Name: James Duncan

Hosting Site: Los Alamos National Laboratory

Mentor: Lori Dauelsberg

Mentor's Signature: Lori Dauelsberg

Abstract: Smuggling of nuclear and radiological material is a complex problem that will need to be analyzed from a variety of angles in order to effectively combat something so rare, yet so hazardous. One crucial tool in the effort to detect and deter smuggling of rad/nuc materials is the radiation portal monitor (RPM)—machines that scan passengers at airports, vehicles and trains at border crossings, and shipping containers at ports throughout the world. While data from RPMs are regularly analyzed at the individual lane level to ensure monitor health and good operation, aggregate analysis of the data has not gone far beyond summary statistics such as number of occupancies and alarm rates. In particular, there are opportunities to carry out interactive exploratory data analysis with the help of tools like the R package Shiny and methods that lend themselves to the visual display of spatiotemporal correlations such as spatial PCA and detrended cross-correlation analysis. By doing so, we discover interesting spatiotemporal patterns in the data that point to specific events or anomalous moments in time that warrant further analysis.

Determining Patterns from Radiation Portal Monitor Data:

Visual and Interactive Exploratory Data Analysis

Introduction

Every year, many gigabytes of data from radiation portal monitors throughout the world are sent back to the US for analysis. In record-setting 2017, Oak Ridge National Laboratory (ORNL) received 533 GB of data from 1023 lanes in 37 countries, amounting to over 100 million occupancies (Oak Ridge National Laboratory, 2018). Such large quantities of data present many challenges, from raw data processing and cleaning to writing efficient algorithms that can take advantage of parallel computing, and finally differentiating signals from the high degree of noise inherent in such a large dataset. Yet, the opportunity to contribute to the mission of non-proliferation and a chance to advance US national security interests—while honing my big data skills—made this project particularly attractive.

Description of the Research Project

What patterns can be discerned from the RPM data, and how might those patterns assist in determining "pattern of life" information from the network of RPMs? This exciting, open-ended question forms the foundation of this research project. So far, few, if any, have attempted to look at this data with the methods of big data including parallel computation, dimensionality reduction, and interactive visualization. The data, therefore, is an untapped resource that may assist national security analysts in the fight against nuclear smuggling.

I worked with a 32 GB subset of 2017 data. From three countries with 42 sites (*Figure 1)* and a total of 243 lanes, including airports, seaports, vehicle and rail crossings at borders throughout the region, this subset contains nearly 6 million occupancies over a two-month period. The raw data is packaged in "daily files", a text file from a single lane with a stream of radiation measurements at varying time increments (*Figure 2*). The 12,854 daily files were preprocessed in parallel to extract data on occupancies (time of day, duration, gamma alarm, neutron alarm, and average speed), detector settings, and detector faults, including gamma high and low, neutron high, RPM cabinet opened a.k.a. tamper, and tamper closed. At around nine hours using four cores, the preprocessing step can be described as painfully slow. However, this is an "embarrassingly parallel" problem in that the upper limit to the benefits of additional parallelization is the number of files being processed.

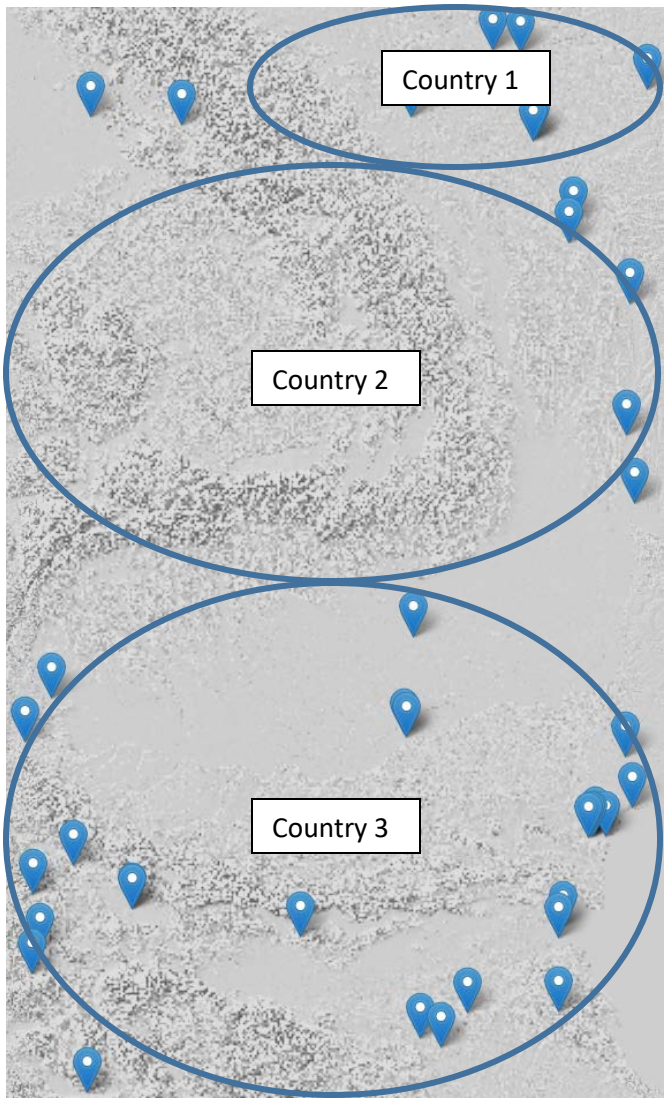

Figure 1. Distribution of sites in the dataset. The map was chosen to keep the countries and sites anonymous. Ovals show approximate positions of the three countries.

Some notable summary statistics of the subset follow. The overall alarm rate was 0.2%, less than the 0.7% rate for the full 2017 dataset. Around 3% of the alarms were neutron alarms, of 69% occurred at a single site. Country 2 had the highest number of occupancies, nearly 60% of the 6 million total. However, more occupancies does not necessarily mean more alarms, as Country 3 had the highest number of alarms and alarm rate (0.4%) while having the second highest number of occupancies. In large part, this is driven by airports, which have heavy traffic but low alarm counts, as naturally occurring radioactive material (NORM) is rare in that setting. Country 2 also has the highest number of faults at about 68% of the 1419 total.

```
GA,000086,000079,000077,000093,03-37-32.745
GS,000080,000088,000100,000086,03-37-33.020
GS,000073,000098,000075,000070,03-37-33.267
GS,000080,000071,000090,000092,03-37-33.365
GS,000096,000084,000074,000088,03-37-33.521
NS,000000,000001,000000,000000,03-37-33.786
GS,000083,000078,000079,000082,03-37-33.882
GS,000073,000069,000088,000072,03-37-34.026
GS,000085,000079,000085,000056,03-37-34.281
GS,000071,000065,000086,000076,03-37-34.376
GS,000073,000076,000071,000064,03-37-34.511
NS,000002,000000,000001,000000,03-37-34.676
GS,000071,000075,000069,000077,03-37-34.778
GS,000087,000077,000082,000072,03-37-34.790
GS,000078,000069,000076,000068,03-37-34.990
GS,000074,000072,000076,000073,03-37-35.110
GS,000079,000067,000077,000083,03-37-35.282
NS,000002,000001,000000,000001,03-37-35.551
GS,000087,000072,000085,000076,03-37-35.565
GS,000074,000079,000095,000079,03-37-35.718
GS,000090,000073,000068,000080,03-37-35.881
GS,000058,000080,000081,000072,03-37-36.125
GS,000088,000083,000076,000069,03-37-36.312
NS,000000,000000,000002,000000,03-37-36.533
GS,000078,000083,000069,000079,03-37-36.543
GS,000093,000075,000068,000096,03-37-36.689
GS,000085,000069,000065,000077,03-37-36.879
GS,000088,000095,000095,000065,03-37-37.144
GS,000075,000073,000086,000073,03-37-37.304
NS,000001,000002,000000,000003,03-37-37.545
GX,000146,004108,000000,000000,03-37-37.561
```

Figure 2. A six-second excerpt from an RPM daily file. From top to bottom, the RPM is in gamma alarm (GA) state, indicating that gamma radiation levels are significantly above the rolling average of background levels. GS and NS are gamma scan and neutron scan, respectively, the default measurements while occupied. At the bottom, GX indicates the occupancy has been cleared.

One of my goals for the project was to build a visual and interactive exploratory data analysis (EDA) tool for the data. To that end, I aggregated the occupancy, alarm, and fault counts for each site at the hourly level. Using the R package Shiny, I plot the hourly counts on an interactive map where the number and type of events

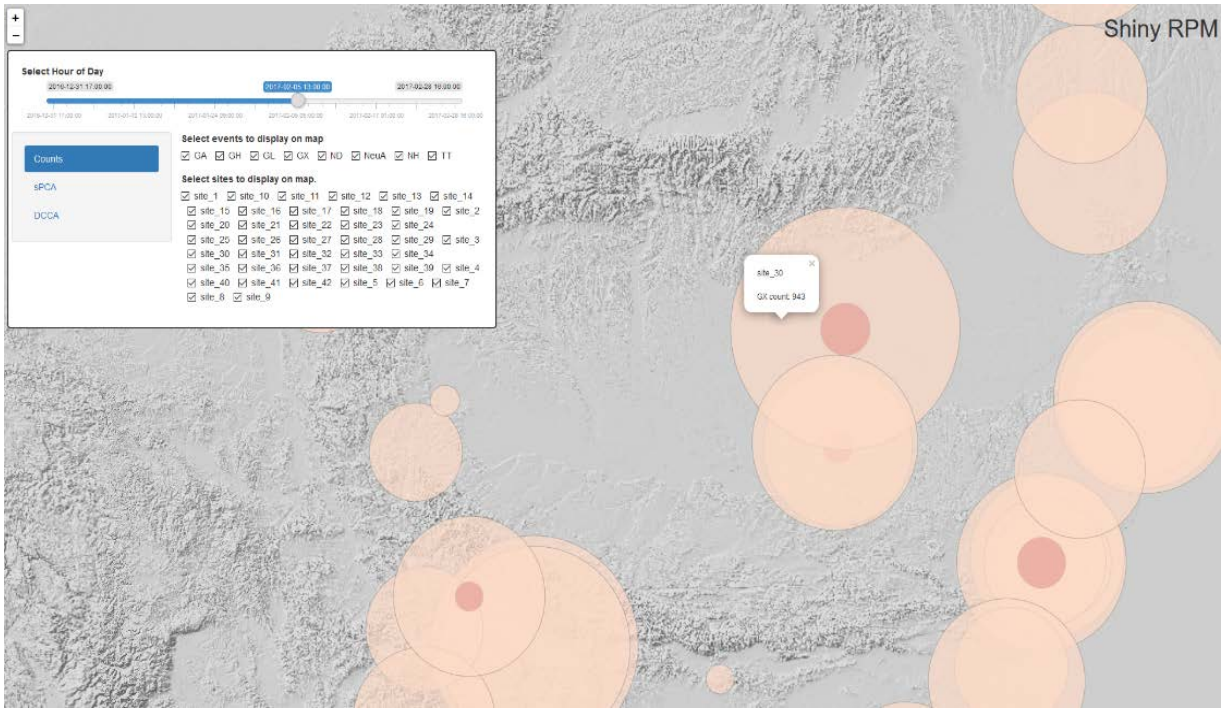

Figure 3. An interactive geographic visualization of the data. A slider allows the user to select the hour to visualize and checkboxes allow the event types and sites to display.

in an hour are represented by the size and color, respectively, of a circle centered at the corresponding site (*Figure 3*). This allows for an informal analysis of the distribution of events over time.

To extract global and local patterns that incorporate both the event counts and information on the spatial distribution of the data, I turn to spatial principal components analysis (sPCA). Let X be the $n \times m$ data matrix, where n is the number of sites and m is the number of variables. In our case, $m = 7$ (GA = gamma alarm, GH = gamma high fault, GL = gamma low fault, GX = occupancy, NA = neutron alarm, NH = neutron high fault, TT = tamper). Traditional PCA seeks to find a number of orthogonal linear combinations Xv (known as scores) of the variables of interest that explain the directions of maximal variability, and so create lower dimensional representations of the data which can reveal hidden clustering. Here, v is an $m \times 1$ vector, which in the case of the first principal component (the direction of maximal variance) is the eigenvector

with largest eigenvalue of the data covariance matrix $X'X$, where $'$ is the matrix transpose operator. Spatial PCA (Jombart, Devillard, Dufour, & Pontier, 2008) incorporates geographic information, encoded in an $n \times n$ connection matrix L, by finding linear combinations of the variables that optimize the product of the data variance and Moran's I, a measure of spatial autocorrelation:

$$C(v) = var(Xv)I(Xv)$$

$$= \frac{1}{n}(Xv)'LXv.$$

Moran's I will be positive when there is a global pattern in the data in which neighbors tend to be more similar to one another, and will be negative when there is a local pattern in the data where neighbors tend to be dissimilar. Thus, to extract both the global and local patterns, we seek the unit vectors v_1 such that $C(v_1)$ is as positive as possible and v_2 such that $C(v_2)$ is as negative as possible. The global and local scores, Xv_1 and Xv_2, are then calculated and visualized.

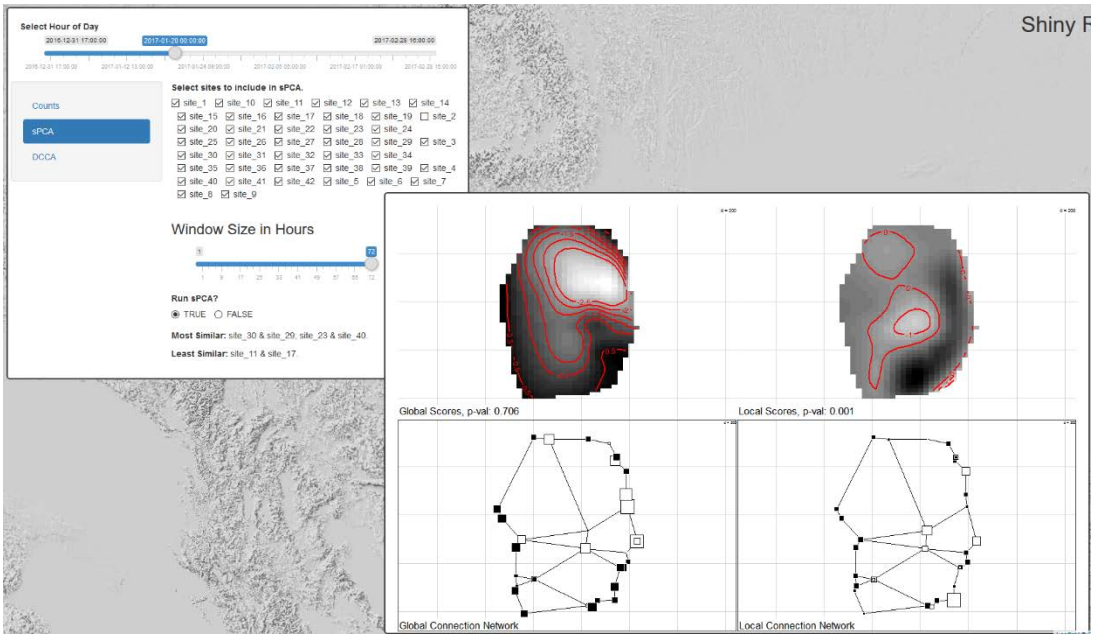

Figure 4. Visualizing sPCA global (left) and local (right) scores. Once again, a slider allows the user to select the hour of interest and checkboxes to select the sites to include in the sPCA. Additionally, a window size greater than 1 can be used to form the data matrix as a weighted average of the counts in the hours preceding and following the selected time. The weights decay exponentially with distance in time from the selected hour. In this case, the window size is 72 hours. Two pairs of the most similar sites are given, as well as the pair of most dissimilar sites.

Figure 4 shows one such visualization. The selected moment is unique in that while the global scores are insignificant ($p = 0.706$), the local scores are significant ($p = 0.001$), indicating that there are significant differences between neighboring sites. Moments of significant global patterns were not uncommon in this dataset. As nearby sites are often similar in some way (e.g., same country, same border, or same traffic type), such global patterns are no surprise, and thus local patterns are of more interest. It should be noted that regardless of whether the patterns are statistically significant or not, this method can give insight into the the structure of the RPM network, revealing the sites that are most similar and dissimilar at a given moment in time.

In this case, a closer look at the local connection network reveals a site in the southeast that appears to be quite distinct from the sites near it (*Figure 5*). Examining the event counts during that time revealed a high number of GL faults at that site. The faults also have an effect on other events at the site, being immediately followed by multiple tampers as operators attempted to address the issue, as well as a corresponding reduction in traffic through the site as the problematic lane was likely closed. Following the TT events, the issue is apparently resolved as there is a large spike in traffic which eventually settles back to normal levels. This example, discovered through casual interaction with the EDA app, demonstrates sPCA's ability to highlight interesting moments in time which can then be investigated at a finer level of detail by an analyst. We obtain a better understanding of the types of challenges and pressures faced by RPM operators. In this case, a high number of unexplained GL faults seems to have seriously impacted operations at the site and led to a large

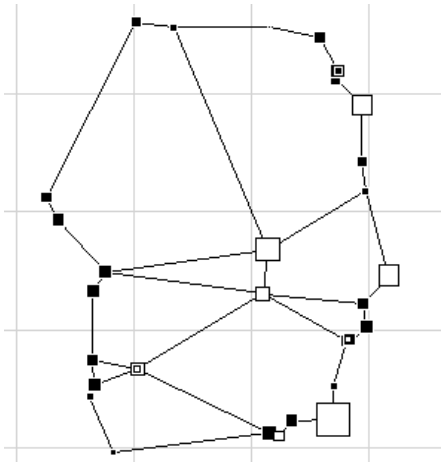

Figure 5. The local scores displayed on the connection network. White for negative scores and black for positive, with square size representing the magnitude of the score. The site represented by a large white square in the southeast of the network is most distinct from its neighbors.

spike in traffic, perhaps an indication that careful secondary screening was relaxed in order to alleviate the backup. Site managers armed with this information can analyze the chain of events with site personnel to ensure that best practices were followed.

A second tool integrated into the Shiny app is temporal evolution of detrended cross-correlation analysis (DCCA). The goal is to single out two sites, filtering by traffic type and direction (entering or exiting the site's country), and see how correlations between the time series of events at each site change over time. When trying to find correlations between two time series, standard cross-correlation analysis has limitations. First, if either series is non-stationary, the assumptions of standard cross-correlation analysis are violated. Second, in standard cross-correlation analysis there is no notion of multi-scale analysis in which correlations are checked at varying lengths of time. Detrended cross-correlation analysis addresses both shortcomings by splitting the time series into overlapping windows of size s and locally fitting a polynomial function to each window, essentially removing any trends that exist on the timescale s. The residuals of these fits are retained and spliced together into a much longer detrended time series, and these residual series are then correlated (Podobnik & Stanley, 2008). The window size s can be varied to examine multi-scale correlations. Moreover, we can examine correlations at specific moments in time by correlating corresponding windows of size s in each of the two series (Yuan, Xoplaki, Zhu, & Luterbacher, 2016).

Figure 6 shows one example of the use of DCCA where the occupancy and gamma alarm time series at one site are correlated. At timescales at or above one day the series tend to be strongly positively correlated. However, during the second quarter of the timeframe examined there is strong negative correlation, and a closer look at the time series reveals that gamma alarms unexpectedly bottom out and then return at a high level in spite of low traffic. Moreover, the alarms

do not track the day night cycle as they do at the beginning and end of the time series. This anomolous moment of time may warrant further investigation. DCCA provides one more tool for the analyst to discover anomalies in the data.

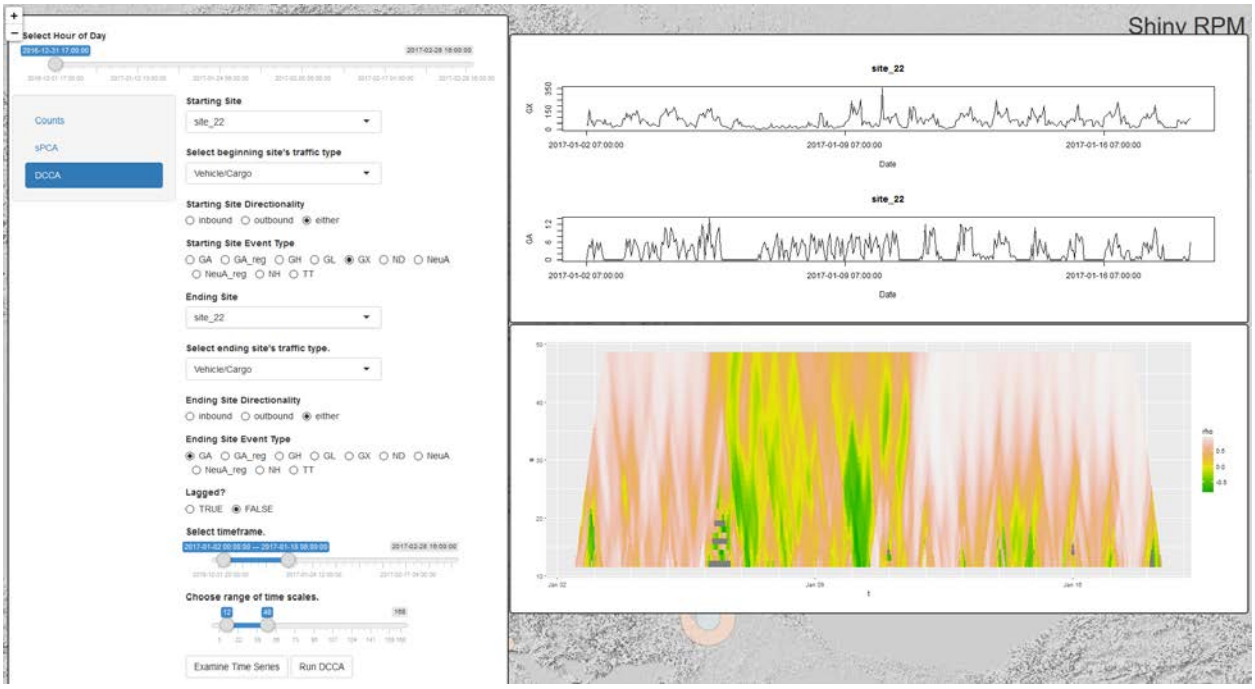

Figure 6. Detrended cross-correlation analysis of occupancy and gamma alarm time series at a single site. The individual time series are shown (upper right), as well as the temporal evolution of DCCA (lower right). In the DCCA plot, the horizontal axis is time and the vertical axis is timescale. White, green, and yellow represent strong positive, strong negative, and no correlation, respectively.

Contributions Made to the Research Project

This project is in its infancy, but I believe that I have demonstrated the potential to use this data to better understand the RPM network, to see the challenges faced by RPM operators and their responses at a high level, and to find anomalous moments in the data that may be relevant to rad/nuc smuggling. I hope that the app I developed will be improved upon or inspire new perspectives on this data.

Skills and Knowledge Gained

Working on this project has further developed my skills in working with big data, including parallel processing, text parsing, data cleaning, and data wrangling. This project presented the opportunity

to work with spatiotemporal data and has developed my data sense in that realm, an area in which I previously had only limited experience. Finally, building Shiny apps is a good skill for any data scientist that regularly uses R to have. I learned a great deal about Shiny and the visual display of geographic and time series data.

Research Experience Impact on Academic and Career Planning

Being at Los Alamos has convinced me that working at the national labs is one of the finest jobs around. I applied to this program because I wanted to get the experience of working at a national lab to see if it would be a good fit, and I have found that indeed it is. Upon graduating in December, I hope to return to the national labs and continue in the mission of non-proliferation.

Relevance to the Mission of DHS

Two important components of the mission of DHS are to "Prevent the unauthorized acquisition, importation, movement, or use of chemical, biological, radiological, and nuclear materials and capabilities within the United States"[1] and "Safeguard and streamline lawful trade and travel."[2] I believe that this research is relevant to both missions in that it demonstrates the power of this data to illuminate operations on the ground and to uncover anomalies that may be relevant to smuggling events. A next step is to take knowledge of known smuggling events and to see how those events manifest themselves in the RPM network, if at all.

Acknowledgements

First and foremost, I'd like to thank my primary mentor Lori Dauelsberg for creating a welcoming environment and guiding me during this research project. Her insight and advice was invaluable at many steps along the way, especially in preparing and facilitating my final presentation. I'd also

[1] https://www.dhs.gov/prevent-terrorism-and-enhance-security
[2] https://www.dhs.gov/secure-and-manage-borders

like to thank my secondary mentor Nate Limback for helping to coordinate the project and secure the data, and for his time and feedback midway through the project. I'm very grateful to Tim Elmont for providing background materials, advice, feedback, and kind words of encouragement. Jennifer Erchinger played an important role in making me feel welcomed in my first week as she spent hours with me giving me an overview of the operating of RPMs, getting me set up with existing software to analyze RPM data, giving me my first tour "behind the fence," and introducing me to many of the people that help ensure these RPMs are effective tools in protecting our national security. John Ambrosiano has been a friendly face throughout the summer and many of his insights helped push me in fruitful directions. Thank you to all the hard-working people at LANL who have made this summer such a fantastic experience, from the student programs coordinators to the employees at various training centers and administrative offices. Finally, thank you to the staff at ORISE and DNDO for this opportunity, especially Denise Young, without whom this summer would not have been possible.

References

Jombart, T., Devillard, S., Dufour, A.-B., & Pontier, D. (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity, 101*, 92-103.

Oak Ridge National Laboratory. (2018). *Nuclear Smuggling Detection and Deterrence FY 2017 Data Analysis Annual Report (ORNL/SPR-2018/25).* Oak Ridge, TN: Oak Ridge National Laboratory.

Podobnik, B., & Stanley, H. E. (2008). Detrended Cross-Correlation Analysis: A new method for analyzing two nonstationary time series. *Phys. Rev. Lett., 100*, 084102.

Yuan, N., Xoplaki, E., Zhu, C., & Luterbacher, J. (2016). A novel way to detect correlations on multi-time scales, with temporal evolution and for multi-variables. *Sci. Rep., 6*, 27707.