SAND2015-5861D

Exceptional service in the national interest

Sandia National Laboratories | Berkeley UNIVERSITY OF CALIFORNIA | SCUREF

Dynamic Analysis Environment (DAE) – An Interactive GUI for Nuclear Forensics

Chad Ummel

Department of Homeland Security – Nuclear Forensics Undergraduate Scholarship Program
University of California, Berkeley – B.A. Physics, 2016
Sandia National Laboratories Org. 1814 – Manager: Amy C. Sun, Mentor: Christopher L. Stork

Special thanks to David S. Stuart and Susan Bodily, the developers of DAE

Abstract

Dynamic Analysis Environment (DAE) is a graphical, user-interactive environment used to facilitate the analysis of data. Current efforts are focused on applying DAE to the analysis of nuclear forensics data. For this purpose, my contribution has been to add the following functionalities to DAE: 1) the K-nearest neighbor supervised learning algorithm, and 2) numerous data-visualization tools to complement the principal component analysis supervised learning method already available in DAE.

Introduction

What is Nuclear Forensics? – Nuclear forensics is the characterization of intercepted nuclear materials to identify evidence of their source and intended use. The Domestic Nuclear Detection Office (DNDO) has supported the creation of national nuclear forensics libraries (NNFLs) of known nuclear materials against which to compare questioned materials as well as the development of multivariate group inclusion/exclusion algorithms to enable the linking of questioned materials/samples to their potential processes or facility of origin. Significant progress has been made in the creation of NNFLs and data analysis tools. However there has been little work done to integrate these data sets and tools into a single package to streamline group exclusion/inclusion analysis. DAE accomplishes this task.

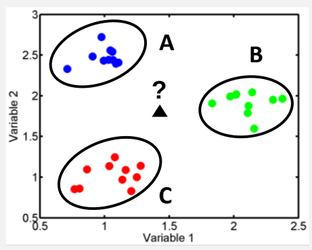
What is group inclusion/exclusion?

Inclusion: Identify a match between signatures of a questioned sample and that of a known group.

Exclusion: Eliminate the possibility that a questioned sample originated from a known group.

Questioned material

Compare signatures for material with known groups

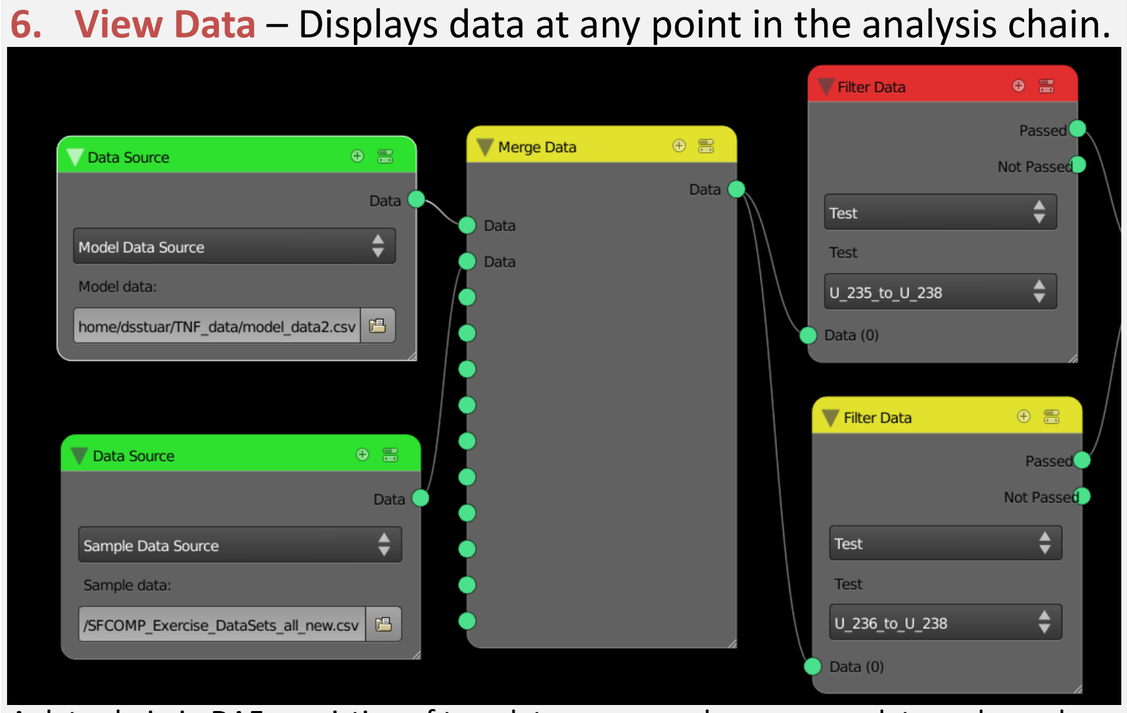


Exclusion - questioned material (?) inconsistent with groups A, B, and C

DAE – Basic Function and Capabilities

DAE is an interactive, configurable environment that facilitates the analysis of very large data sets though the use of predefined analytical modules, called nodes. These nodes can be arranged into an analytical "chain" of data processing steps. Available nodes include:

1. **Data Source** – Extracts known and unknown data from files.
2. **Merge Data** – Combines data streams.
3. **Filter Data** – Reduces data based on defined parameters (i.e. reactors or specific isotopic measurements).
4. **Augment Data** – Prepares PCA or KNN model based on known/training data.
5. **Relate Data** – Relates test data to PCA or KNN model.
6. **View Data** – Displays data at any point in the analysis chain.



A data chain in DAE consisting of two data source nodes, a merge data node, and two filter data nodes. Green nodes have completed their functions, yellow nodes are in the process of completing their functions, and red nodes have yet to begin.

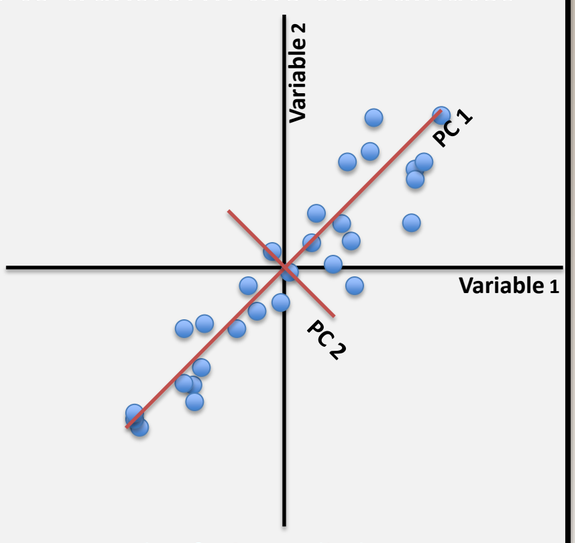
Nuclear Forensics Data – SFCOMPO

In constructing and evaluating the performance of the PCA and KNN group inclusion/exclusion methods, known nuclear material data were taken from the open-source Spent Fuel Isotopic Composition (SFCOMPO) database, which consists of isotopic measurements of spent fuel samples from fourteen different nuclear reactors from around the world.

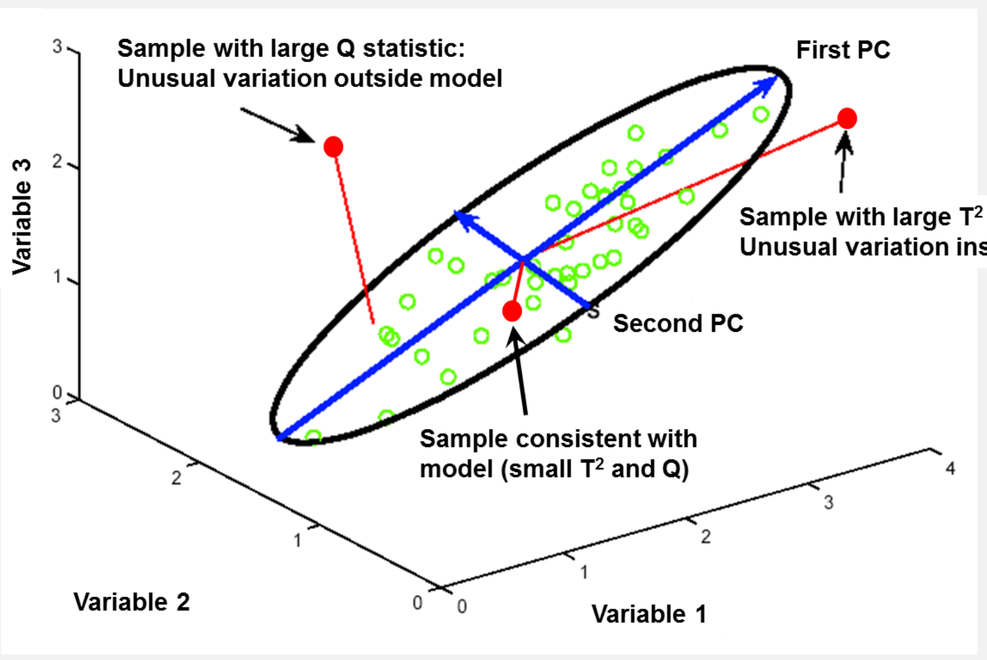
This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2012-DN-130-NF0001-02. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

Principal Component Analysis (PCA)

Theory – The objective of PCA is to transform the coordinates of a data matrix to a new set of axes, called principal components, which optimally describe the data variance. Principal components (PCs) are assigned to the data matrix such that the first principal component explains the maximum amount of variation possible in the data set in one direction. Successive PCs are defined such that they are orthogonal to the previous PCs and describe the maximum amount of remaining variation in the data. Often, a small number of PCs describe a large percentage of the variance in the data, and subsequent PCs may be ignored. Once PCA is performed on model data, sample data can be compared to the model data by projecting the measurement variables of each sample into the PC space. Sufficiently small values of the *Q*, Hoteling's T^2, and Hawkins' T_H^2 statistics indicate the sample is consistent with the model data.



An example of PCA applied to a two-dimensional data set. The first PC accounts for the maximum variance in the data set in one direction, the second the maximum remaining variation.



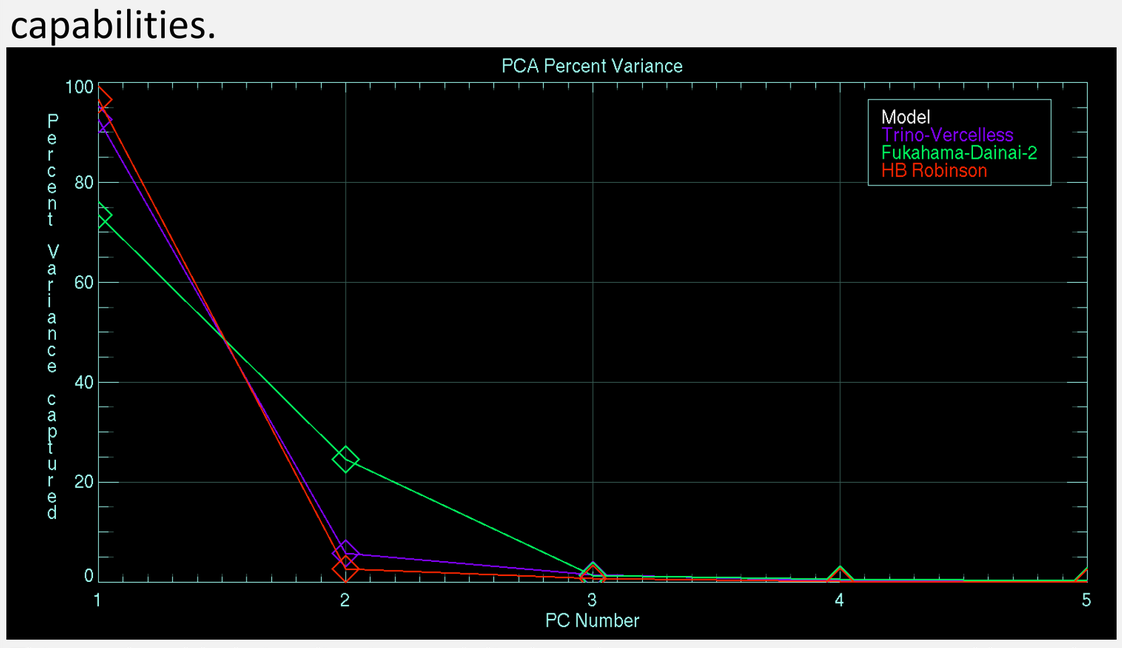
An example of comparing sample data to a PC model using *Q* and T^2 values.

PCA Sample To Model Correlation - Qn

Model	RF	1	2	3	4	5	6	7	8	9	10	11
Fukahama-Dainai-2	1	0.6354	0.3976	0.2017	0.5457	0.5739	0.0966	0.4430	0.2466	0.0444	0.0304	0.1993
HB Robinson	1	0.4286	0.6056	0.1934	0.4459	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Trino-Vercelless	1	0.0054	0.0000	0.0000	0.0003	0.4246	0.7490	0.3026	0.4599	0.6397	0.2493	0.3147

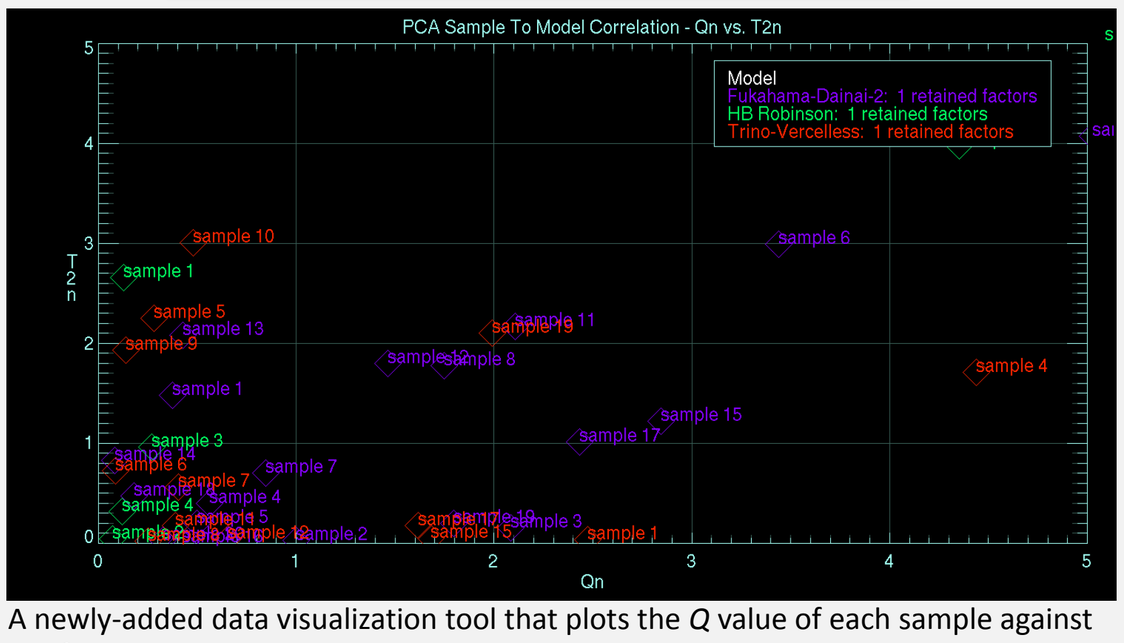
A PCA Sample to Model table output by DAE. The cells give each of 11 samples' *Q-probability* to PC models formed by data from three reactors. Higher *Q*-probabilities indicate a close match with a model and are designated by brighter colors.

Additions to DAE – A number of data visualization tools have been added to DAE to complement the existing PCA capabilities.



This newly-added visualization tool displays the percent variance captured by each of the five PCs for models formed using data from three reactors. This feature allows the user to determine how many PCs need to be retained in order to adequately model the data. An additional tool displays both the percent variance captured by each PC and the total variance captured up to that point by previous PCs in a bar graph.

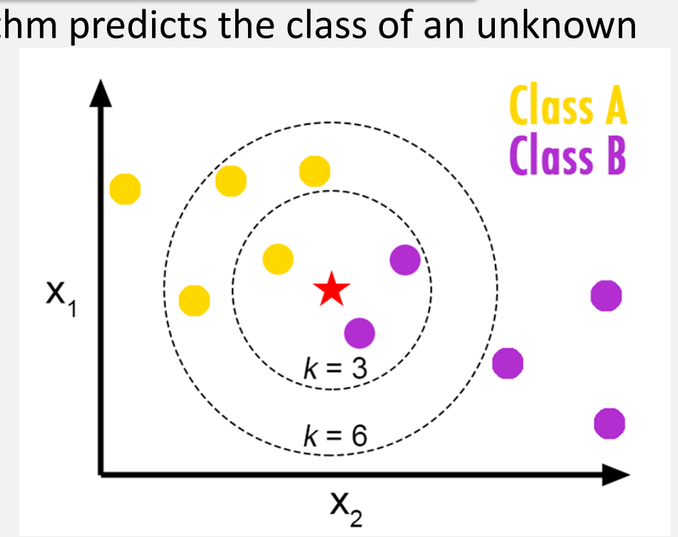
PCA (continued)



A newly-added data visualization tool that plots the *Q* value of each sample against its T^2 value, in this case for three different PCA reactor models.

K-Nearest Neighbors (KNN)

Theory – The KNN algorithm predicts the class of an unknown sample based on the majority vote of the known classes of the *k* sample(s) nearest to it. Each sample is then assigned a *goodness value*. A low goodness value indicates that a sample is well-contained by known samples in the class to which it is assigned. If a sample has a goodness value that is higher than the *goodness value threshold* of its assigned class, the sample is deemed to not be a member of that assigned class.



An example of KNN performed on an unknown sample (red star). If *k* = 3, the majority of the samples' 3 neighbors belong to Class B and the sample is assigned to Class B. If *k* = 6, the sample is assigned to Class A.

Addition to DAE – KNN is now available in DAE as a method of group inclusion/exclusion. A "KNN Training" augment data node allows users to choose a distance metric (i.e. Euclidean, City Block, Minkowski, etc.) for which to compute the *k* nearest neighbors of sample data points. The KNN Training augment node both determines the goodness value threshold of each individual class as well as performs the KNN algorithm on each class, recording the percentage of known samples correctly classified for each value of *k*, thus allowing the user to choose an optimal value of *k* to use in subsequent calculations. A "KNN Sample to Training" relate data node assigns a class and goodness value to each unknown sample. If the goodness value of a sample exceeds the threshold of the class to which it is assigned, the assignment is removed.

KNN Sample To Training Classification

Training Class	K	1	2	3	4	5	6	7	8	9
Fukahama-Dainai-2	3	NaN	0.520439	-1.20007	-1.19560	0.197979	NaN	NaN	0.0860031	0.466796
Calvert Cliffs	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
HB Robinson	3	-1.65620	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

The KNN Sample to Training classification table output by DAE. The cells give each of 9 samples' goodness values for classes (formed by data from three reactors) to which they are assigned for *k* = 3. Lower (or negative) goodness values indicate a close match with a class and are designated by brighter colors. A value of NaN in a cell indicates that a sample has not been assigned to the corresponding class.

Future Plans

Data visualization tools to complement the new KNN group inclusion/exclusion method will be added.

Work will also be performed to evaluate and compare the success with which the PCA and KNN algorithms in DAE relate questioned data samples to reactors in the SFCOMPO database.

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2012-DN-130-NF0001-02. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.