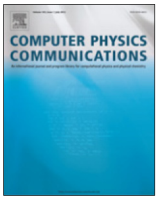
Computer Physics Communications 210 (2017) 60–71

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc



Dynamic analysis environment for nuclear forensic analyses

C.L. Stork[a], C.C. Ummel[b,*], D.S. Stuart[a], S. Bodily[a], B.L. Goldblum[b]

[a] Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185, USA
[b] University of California, Berkeley, CA 94720, USA

ARTICLE INFO

Article history:
Received 1 June 2016
Received in revised form
7 September 2016
Accepted 26 September 2016
Available online 5 October 2016

Keywords:
Nuclear forensics
Principal component analysis
k nearest neighbors algorithm

ABSTRACT

A Dynamic Analysis Environment (DAE) software package is introduced to facilitate group inclusion/exclusion method testing, evaluation and comparison for pre-detonation nuclear forensics applications. Employing DAE, the multivariate signatures of a questioned material can be compared to the signatures for different, known groups, enabling the linking of the questioned material to its potential process, location, or fabrication facility. Advantages of using DAE for group inclusion/exclusion include built-in query tools for retrieving data of interest from a database, the recording and documentation of all analysis steps, a clear visualization of the analysis steps intelligible to a non-expert, and the ability to integrate analysis tools developed in different programming languages. Two group inclusion/exclusion methods are implemented in DAE: principal component analysis, a parametric feature extraction method, and *k* nearest neighbors, a nonparametric pattern recognition method. Spent Fuel Isotopic Composition (SFCOMPO), an open source international database of isotopic compositions for spent nuclear fuels (SNF) from 14 reactors, is used to construct PCA and KNN models for known reactor groups, and 20 simulated SNF samples are utilized in evaluating the performance of these group inclusion/exclusion models. For all 20 simulated samples, PCA in conjunction with the *Q* statistic correctly excludes a large percentage of reactor groups and correctly includes the true reactor of origination. Employing KNN, 14 of the 20 simulated samples are classified to their true reactor of origination.

1. Introduction

Nuclear forensics is a branch of science in which questioned nuclear materials are characterized with regard to their isotopic and elemental composition, age, physical state, history, and provenance [1]. The characterization and interpretation of a questioned nuclear material may require the integration of information from a wide array of sources, including visual inspection and laboratory analyses of the material, computer modeling, and a comparison of the features or signatures of the questioned nuclear material with those of known materials [2].

Nuclear forensics data sets or libraries of known nuclear materials have been developed against which to compare questioned materials. These libraries, in conjunction with multivariate pattern recognition algorithms, enable the linking of questioned materials to their potential processes, location, or fabrication facility [3–12]. This linking procedure is performed by systematically comparing the multivariate features/signatures (e.g., isotopic or trace element measurements) for a questioned nuclear material to the signatures for materials originating from different, known groups or classes (e.g., specific nuclear reactors, processes, or locations), enabling both group exclusion and inclusion. Exclusion refers to the process of eliminating the possibility that a questioned material originated from a particular group, based on a rigorous statistical comparison of the signatures for the questioned material and known materials from that group. In contrast, inclusion refers to the process of identifying a statistically significant match between the signatures for the questioned material and the known materials from a group, indicating that the questioned material may have originated from this group.

To date, these nuclear forensics data sets and analysis tools have largely been developed independently, with minimal coordination and no formal plan for eventual integration. Accordingly, a data analyst must develop his or her own ad hoc, labor-intensive approach for tying together the various data sets and analysis tools to perform a group inclusion/exclusion analysis. The objective of this work is to integrate the independently developed nuclear forensics data sets and data analysis tools into a graphical, user-interactive test bed to facilitate group inclusion/exclusion method testing, evaluation and comparison using Sandia National

[*] Corresponding author.
E-mail address: chadummel@berkeley.edu (C.C. Ummel).

http://dx.doi.org/10.1016/j.cpc.2016.09.019
0010-4655/© 2016 Elsevier B.V. All rights reserved.

Laboratories' Dynamic Analysis Environment (DAE) software package. DAE, which has been under active development since 2010, provides a highly interactive and configurable computational environment that facilitates the analysis of very large data sets through the use of predefined analytical modules. DAE is designed to allow the incorporation of new algorithms by wrapping them into modules with simple and standardized interfaces to the source data and the data exchanged with other modules. DAE supports modules developed in C/C++, Fortran, Python, MATLAB, IDL, and PERL, eliminating the need to port algorithms from one language to another. In DAE, the user can tailor the analysis by creating analytical flows (networks) using a palette of available data retrieval, data manipulation, and data visualization modules. The interface is "drag and drop" and provides immediate feedback of processing status as well as complete access to all intermediate and final results. This feedback is available at all times, even during the assembly of the analysis network itself. Advantages of using DAE in the development of a test bed for group inclusion/exclusion include built-in query tools for retrieving data of interest from a database, the recording and documentation of all analysis steps, a clear visualization of the analysis steps intelligible to a non-expert, and the ability to integrate analysis tools developed in different programming languages.

This manuscript documents the development of a DAE-based software package for nuclear forensics data visualization and group inclusion/exclusion analysis. Section 2 provides an overview of the two group inclusion/exclusion methods, namely principal component analysis (PCA) and k nearest neighbors (KNN), that have been integrated into DAE. Section 3 describes Spent Fuel Isotopic Composition (SFCOMPO), an open source international database of isotopic compositions for spent nuclear fuels (SNF) [13], used in constructing PCA and KNN models, and simulated SNF samples employed in evaluating the performance of these group inclusion/exclusion models. Section 4 describes the framework of the DAE software package. Section 5 presents the group inclusion/exclusion results obtained in applying the DAE software package to SFCOMPO and the simulated SNF samples. Section 6 offers some concluding remarks.

2. Methodology

Throughout this section, scalars are represented by italics, e.g., n. Column vectors are denoted by boldface lowercase letters, e.g., $\mathbf{x}$. Matrices are represented by boldface uppercase letters, e.g., $\mathbf{X}$. The transpose of a matrix or vector is symbolized by a superscripted 'T', e.g., $\mathbf{X}^{\mathrm{T}}$ and $\mathbf{x}^{\mathrm{T}}$.

2.1. Principal component analysis

The objective of PCA is to extract a new set of features that optimally describe the variation or major trends within a given data matrix $\mathbf{X}$, composed of m samples or materials and n variables [14]. The matrix $\mathbf{X}$ consists of measurements representative of a group of interest, such as isotopic measurements for SNF materials from a reactor. Typically, the PCA modeling procedure begins by preprocessing the matrix $\mathbf{X}$ using either mean centering, autoscaling, or range scaling [15]. For a matrix $\mathbf{X}$ of inherent linear dimensionality l, with $l \le \min\{m, n\}$, PCA decomposes $\mathbf{X}$ into a set of l rank 1 matrices, arranged in order of decreasing eigenvalue, plus a residual matrix $\widetilde{\mathbf{X}}$, corresponding to noise or information irrelevant to describing the groups of interest: [16]

$$\begin{aligned}\mathbf{X} &= \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \cdots + \mathbf{t}_l\mathbf{p}_l^{\mathrm{T}} + \widetilde{\mathbf{X}} \\ &= \mathbf{T}_l\mathbf{P}_l^{\mathrm{T}} + \widetilde{\mathbf{T}}\widetilde{\mathbf{P}}^{\mathrm{T}} \\ &= \bar{\mathbf{T}}\bar{\mathbf{P}}^{\mathrm{T}} \end{aligned} \tag{1}$$

where $\widetilde{\mathbf{X}} = \widetilde{\mathbf{T}}\widetilde{\mathbf{P}}^{\mathrm{T}}$, $\bar{\mathbf{T}} = [\mathbf{T}_l\ \widetilde{\mathbf{T}}]$, and $\bar{\mathbf{P}} = [\mathbf{P}_l\ \widetilde{\mathbf{P}}]$. The score vector, $\mathbf{t}_i$, can be interpreted as the samples' coordinates for principal component i as defined by the new basis or loading vector, $\mathbf{p}_i$. The principal component subspace is modeled by the span of $\mathbf{P}_l$, while the residual subspace is modeled by the span of $\widetilde{\mathbf{P}}$. In the context of modeling isotopic measurements acquired for a set of SNF materials from a nuclear reactor, $\mathbf{P}_l$ models correlations in that reactor among the measured isotopes.

Similarly, the covariance matrix $\mathbf{S}$ can be decomposed using eigenanalysis:

$$\begin{aligned}\mathbf{S} &= \frac{1}{m-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} \\ &= \bar{\mathbf{P}}\bar{\boldsymbol{\Lambda}}\bar{\mathbf{P}}^{\mathrm{T}}, \end{aligned} \tag{2}$$

where

$$\begin{aligned}\bar{\boldsymbol{\Lambda}} &= \frac{1}{m-1}\bar{\mathbf{T}}^{\mathrm{T}}\bar{\mathbf{T}} \\ &= \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}. \end{aligned} \tag{3}$$

Here, $\bar{\boldsymbol{\Lambda}}$ is the n by n diagonal eigenvalue matrix that contains the eigenvalues of the covariance matrix $\mathbf{S}$ in descending order.

PCA can be used as a form of unsupervised pattern recognition to recognize intrinsic groups of samples within a given data matrix, or as a supervised pattern recognition method where a separate model is developed for each group. The procedure for using PCA as a supervised pattern recognition method is now summarized. Using the PCA model developed for the matrix $\mathbf{X}$, various statistics, such as the Q statistic, Hotelling's T^2 statistic, and Hawkins' T_H^2 statistic, can be calculated for an appropriately preprocessed questioned sample vector, $\mathbf{x}$ (n by 1), as tests for group inclusion and exclusion [14].

The Q statistic provides a quantitative measure of how far a questioned sample vector lies outside the principal component subspace [16]:

$$Q = \mathbf{x}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P}_l\mathbf{P}_l^{\mathrm{T}})\mathbf{x} \tag{4}$$

where $\mathbf{I}_n$ is an n by n identity matrix. The questioned sample vector, $\mathbf{x}$, is considered consistent with the modeled group (e.g., making a case for group inclusion) if $Q \le Q_\alpha$, where Q_α corresponds to the upper control limit for the Q statistic at a significance level α. Likewise, if $Q > Q_\alpha$, this makes a strong case for excluding the questioned sample from potential membership in the modeled group. The upper control limit, Q_α can be calculated using the following expression [14,16]:

$$Q_\alpha = \theta_1\left(\frac{c_\alpha\sqrt{2\theta_2h_0^2}}{\theta_1} + 1 + \frac{\theta_2h_0(h_0-1)}{\theta_1^2}\right)^{\frac{1}{h_0}} \tag{5}$$

where

$$\theta_i = \sum_{j=l+1}^{n}\lambda_j^i, i = 1, 2, 3 \tag{6}$$

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2} \tag{7}$$

l is the number of principal components retained in the model, n is the total number of principal components, and c_α is the normal deviate corresponding to the upper $(1-\alpha)$ percentile. It is noted that the upper control limit, Q_α, is a function of the eigenvalues of the modeled group data. Eq. (5), expressing the critical value of Q given α, can be inverted to obtain the normal deviate, c, corresponding to the probability associated with a given Q value:

$$c = \theta_1\frac{\left[\left(\frac{Q}{\theta_1}\right)^{h_0} - \frac{\theta_2h_0(h_0-1)}{\theta_1^2} - 1\right]}{\sqrt{2\theta_2h_0^2}}. \tag{8}$$

In this way, the Q value for a questioned sample can be converted to the corresponding probability value.

In contrast to the Q statistic, Hotelling's T^2 statistic provides a measure of the variation of a questioned sample vector within the principal component subspace [14]:

$$T^2 = \mathbf{x}^{\mathrm{T}}\mathbf{P}_l\boldsymbol{\Lambda}_l^{-1}\mathbf{P}_l^{\mathrm{T}}\mathbf{x}. \tag{9}$$

Under the condition that the data follow a multivariate normal distribution, a statistical confidence limit for the T^2 values can be calculated employing the equation:

$$T_\alpha^2 = \frac{l\,(m-1)}{m-l}F_{l,\ m-l,\ \alpha} \tag{10}$$

where m is the number of samples utilized to develop the PCA model, l is the number of principal components retained in the model, and $F_{l,m-l,\alpha}$ is an F distribution with l and $m-l$ degrees of freedom at significance level α. For a specified significance level α, the questioned sample vector is considered consistent with the modeled group if $T^2 \le T_\alpha^2$, thereby making a case for group inclusion. In contrast, if $T^2 > T_\alpha^2$, this makes a strong case for group exclusion. The T^2 value for a questioned sample can also be converted to the corresponding probability value.

Hawkins' T_H^2 statistic is a parallel calculation of Hotelling's T^2 statistic in the residual subspace: [17]

$$T_H^2 = \mathbf{x}^{\mathrm{T}}\widetilde{\mathbf{P}}\widetilde{\boldsymbol{\Lambda}}^{-1}\widetilde{\mathbf{P}}^{\mathrm{T}}\mathbf{x}. \tag{11}$$

One liability of Hawkins' T_H^2 statistic compared to the Q statistic is that inversion of the residual eigenvalue matrix is required, which can result in numerical errors when some of the residual eigenvalues are nearly zero. T_H^2 has a T^2 distribution like that represented in Eq. (10), with the exception that l is replaced by $(n-l)$ [14].

2.2. *k* nearest neighbors algorithm

The k nearest neighbors algorithm classifies a questioned sample according to the majority vote of its k nearest neighbors in the training set in multidimensional space [15]. In the event of a tie, the closer neighbors are assigned a larger weight. Nearness is measured using a distance metric such as the Euclidean distance. The Euclidean distance, d, between two samples $\mathbf{x}$ and $\mathbf{y}$ is calculated in n dimensions as

$$d = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}. \tag{12}$$

To classify a questioned sample, the distances are calculated between the questioned sample and training samples of known group membership. The k closest training samples are employed to classify the questioned sample. The questioned sample is assigned to the group having the most members among the k nearest neighbors.

In the traditional implementation of KNN, each questioned sample is assigned to one of the known groups represented by the training set. Thus, the training set is assumed to represent the full spectrum of groups anticipated in the environment. This assumption is problematic in that the questioned sample may not belong to any of the modeled groups [18]. In KNN, a "goodness value" criterion has been proposed to account for the case where a questioned sample is not a member of any of the modeled groups [15]. The goodness value criterion involves validating the group prediction for the questioned sample by comparing the distance from the questioned sample to its predicted group relative to an expected distance for known members of that group. The first step in implementing the goodness value criterion is to calculate the distance, d_{unk}, from the questioned sample to its nearest neighbor in the group predicted by KNN. Next, this calculated distance, d_{unk}, is compared to the interpoint one nearest neighbor distances for each of the training samples that are members of the predicted group, g. The mean, $\overline{d_g}$, and standard deviation, $s(d_g)$, of these training set nearest neighbor distances are calculated, and a goodness value, G, is computed as

$$G = \frac{d_{unk}-\overline{d_g}}{s(d_g)}. \tag{13}$$

This goodness value is indicative of the number of standard deviation units the distance of the questioned sample is from the mean group distance. Confidence in the group assignment of the questioned sample increases as the goodness value decreases. In practice, the questioned sample can be excluded from membership in its predicted group if its goodness value exceeds the maximum goodness value for the training samples that are a member of that group. It is emphasized that while the goodness value criterion can be used to allow for group exclusion with KNN, no specific statements can be made regarding the statistical confidence of this group exclusion assessment.

3. SFCOMPO database and simulated samples

SFCOMPO is an open-source international database of isotopic compositions for SNF obtained through post-irradiation experiments [13]. SFCOMPO consists of SNF isotopic compositions for 14 nuclear reactors in 4 countries, namely Germany, Italy, Japan, and the United States. The SFCOMPO database was originally compiled by the Japan Atomic Energy Research Institute (JAERI), and in 2002 this database was transferred to the Organization for Economic Cooperation and Development/Nuclear Energy Agency (OECD/NEA).

Table 1 provides a summary of the SFCOMPO database. SFCOMPO compiles measured isotopic compositions from 14 reactors, including 7 pressurized water reactors (PWRs) and 7 boiling water reactors (BWRs). In total, data are available from 246 SNF samples. Measured isotopes contained within the SFCOMPO database include U, Pu, Am, Cm, and several fission products (e.g., Nd, Cs, and Sr). As SFCOMPO compiles data from the open literature originally acquired by laboratories in different countries employing slightly different protocols, not all of the isotopes are available for each of the 14 reactors. In fact, only U and Pu isotopic compositions are available for all 14 reactors.

As detailed in Section 5, PCA and KNN models were constructed employing DAE using the SNF isotopic composition data for individual reactors and combined sets of related reactors. In certain instances, data for related reactors were combined due to the limited number of samples available for some reactors, such as Genkai-1 and H.B. Robinson Unit 2. The set of 5 isotopic ratios, namely ^{235}U/^{238}U, ^{236}U/^{238}U, ^{240}Pu/^{239}Pu, ^{241}Pu/^{239}Pu, and ^{242}Pu/^{239}Pu, available for all 14 reactors within the SFCOMPO database were used in developing the PCA and KNN models.

Simulated SNF sample specifications were generated by Argonne National Laboratory (ANL) and treated as questioned or test samples. Five sets of simulated samples were created, with each set consisting of four SNF samples originating from one of the 14 reactors represented in the SFCOMPO database. Thereby, in total, there are 20 simulated SNF samples. Isotopes included in the simulated sample specifications are ^{234}U, ^{235}U, ^{236}U, ^{238}U, ^{238}Pu, ^{239}Pu, ^{240}Pu, ^{241}Pu, ^{242}Pu, ^{241}Am, ^{242}Am, ^{237}Np, ^{137}Cs, and ^{99}Tc. As detailed in Section 5, the 20 simulated SNF samples were compared in DAE to the PCA and KNN models constructed for individual or related sets of reactors, enabling group inclusion/exclusion analysis of these

Table 1
Summary of SFCOMPO database and true reactor of origin for simulated SNF samples.

Label	Reactor	Country	Reactor type	Number of SNF samples	Simulated SNF samples
A	Calvert Cliffs No. 1	United States	PWR	9	
B	Cooper	United States	BWR	6	
C	Fukushima-Daiichi-3	Japan	BWR	36	
D	Fukushima-Daini-2	Japan	BWR	18	Samples 13–20
E	Genkai-1	Japan	PWR	2	
F	Gundremmingen	Germany	BWR	12	
G	H.B. Robinson Unit 2	United States	PWR	6	Samples 1–4
H	JDPR	Japan	BWR	30	
I	Mihama-3	Japan	PWR	9	
J	Monticello	United States	BWR	30	
K	Obrigheim	Germany	PWR	23	
L	Takahama-3	Japan	PWR	16	
M	Trino Vercellese	Italy	PWR	39	Samples 5–12
N	Tsuruga-1	Japan	BWR	10	

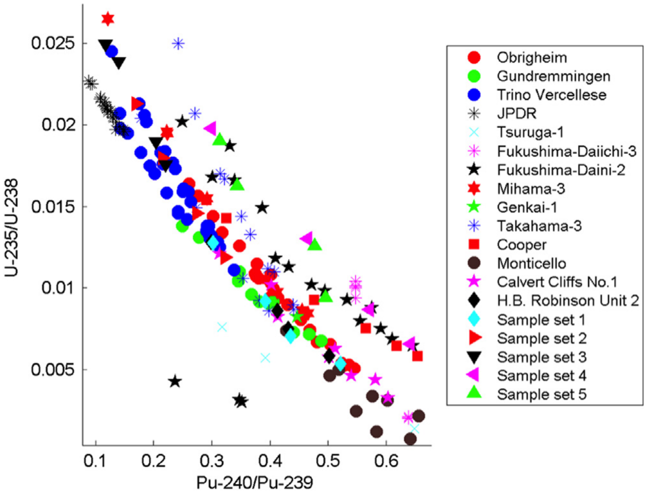


Fig. 1. Bivariate isotopic ratio plot for ^{235}U/^{238}U versus ^{240}Pu/^{239}Pu.

questioned samples. The rightmost column of Table 1 lists the true reactor of origination for each of the 20 simulated SNF samples.

Fig. 1 presents the bivariate isotopic ratio plot for ^{235}U/^{238}U versus ^{240}Pu/^{239}Pu for the samples originating from the 14 reactors in the SFCOMPO database and the 20 simulated SNF samples. Sample set 1 overlaps with samples from the H.B. Robinson Unit 2, Gundremmingen, and Trino Vercellese reactors. Sample sets 2 and 3 bear a resemblance to the Trino Vercellese and Mihama-3 samples in terms of their positions in Fig. 1. Sample sets 4 and 5 correspond with the positions of the Fukushima-Daini-2 samples. With regard to outliers, three Fukushima-Daini-2 samples, localized in the lower left section of Fig. 1, deviate significantly from the other samples for this reactor.

4. Dynamic analysis environment framework

The Dynamic Analysis Environment (DAE) is an extension of the Blender open source software package [19]. The native Blender application is designed for creating animated videos and does not, in and of itself, include any data analysis capabilities. Blender does, however, employ a number of features that make it attractive for extension into the world of data analytics. These features exist primarily within the node compositor panel in Blender. Node compositing is used to apply finishing touches on a frame-by-frame basis as output animation is constructed. The bulk of the native Blender node compositing features deal with image processing, and are accessed via a variety of host nodes. These features include (1) interactive modular network design, (2) advanced image enhancement algorithms, (3) multi-threaded processing across parallel network paths, (4) node-to-node data exchange with unlimited branching, (5) video compilation capability with frame-variable (parametric) nodal inputs, and (6) real-time network processing adaptation in response to user inputs.

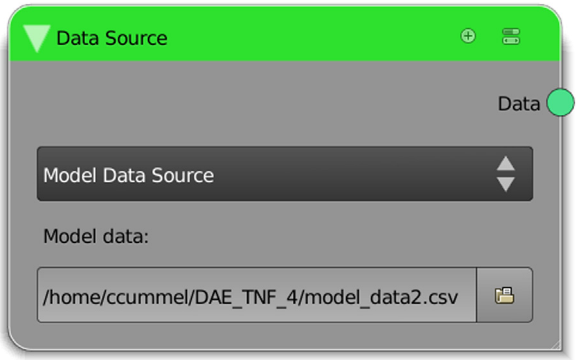
The DAE extension adds a number of capabilities required for a data analysis application. These include (1) an integrated SQLITE3 database [20] and thread locking to store all source, intermediate and final data, (2) real-time node progress feedback, (3) embedded knowledge of data relationships, (4) graph theory algorithms for tracing all data relationships implied by the user-defined nodal networks, (5) eight new node types for hosting data analysis functionality, (6) expanded information channel exchange between connected source and destination nodes, and (7) new socket types for exchanging data and date/time information.

Output information from data analysis networks is accomplished by utilizing existing Blender nodes for directing output to the screen, to still frame files (e.g., PNG), or to composite animations of multiple frames (e.g., AVI). Animations can be made using parametrically-driven node settings.

Eight new data analysis nodes were added to the Blender package, including Data Source (Section 4.1), Augment Data (Section 4.2), Filter Data (Section 4.3), Merge Data (Section 4.4), Relate Data (Section 4.5), View Data (Section 4.6), Compose View (Section 4.7), and Date/Time (Section 4.8). In a general sense, these basic functions form the basis for any data analysis, and encompass the basic steps of gathering the input data, extending the input data using mathematical algorithms, relating sets of initially independent input data to each other, reducing the volume of data with filtering, and finally, visualization of the results. Each of these node types is described below along with the common PCA/KNN modules, PCA-specific modules, and KNN-specific modules incorporated within each node type. Additional data analysis or visualization codes written in C/C++, FORTRAN, MATLAB, IDL, Python, and PERL can be wrapped into a corresponding node type with simple and standardized interfaces to the source data and the analyzed data exchanged with other nodes. Sections 4.9 and 4.10 provide examples of how the nodes can be linked to perform PCA and KNN group inclusion/exclusion analysis using training data from the SFCOMPO database and test data from the simulated samples. The eight new data analysis nodes are described in detail below.

4.1. Data source node

The Data Source node, shown in Fig. 2, is used to import data into the internal database and represents the beginning of any analysis network. Analytic modules are available via dropdown

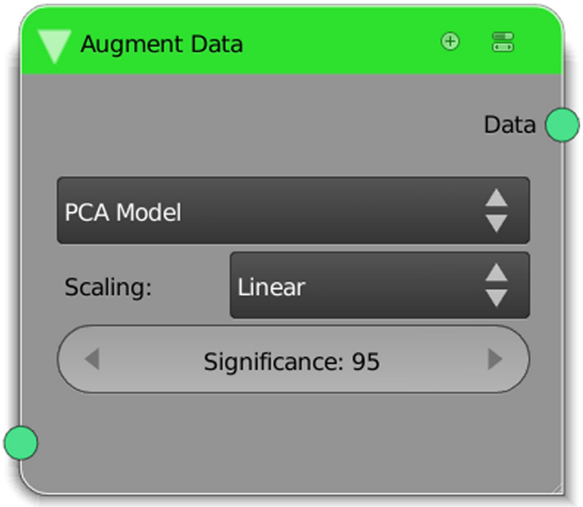


Model Data Source

Model data:

/home/ccummel/DAE_TNF_4/model_data2.csv

Fig. 2. Data Source node. The Model Data Source and Sample Data Source modules can be chosen via a dropdown menu. Note the single output socket on the right side of the node.



PCA Model

Scaling: Linear

Significance: 95

Fig. 3. Augment Data node with the PCA Model module selected. The model can be computed with linear or logarithmic scaling, selectable via the second dropdown menu, and with a significance that can also be chosen using a value input control. Other modules can be selected using the top dropdown menu.

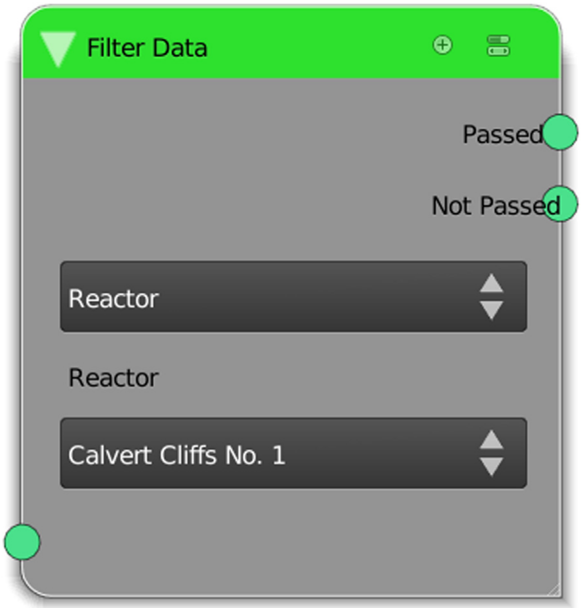
menus to perform the import of source data into one or more internal database tables. Multiple data source nodes may be used within a network.

The common PCA/KNN modules are:

- **Model Data Source:** This module loads the SFCOMPO modeling/training data into a series of related tables within the internal database. Tables include reactors, assemblies, fuel rods, fuel pins, labs, tests, and results along with the associated set of relationship tables.
- **Sample Data Source:** This module loads the questioned sample data into a series of related tables within the internal database. Tables include samples, labs, tests and results along the associated set of relationship tables.

4.2. Augment data node

The Augment Data node, shown in Fig. 3, is used to extend the source data by invoking analytic processes via dropdown selections. During augmenter processing, one or more additional tables are created within the database to store the processing results. Relationships between new table information and the table information used by the augmenter are established and maintained within the database. The Augment Data nodes have a single input socket and a single output socket. Input sockets provide the table names of all locally relevant database tables for use in the augmenter processing. Output sockets contain the same information, plus the names of any additional tables created during the associated node processing. All table names are prepended by the name of the augmenter instance that created them. As a result, there can be any number of instances of any one node type and its associated table(s). This allows parallel execution paths involving the same type of augmenter (or any other node type), without the chance of commingling the data.



Reactor

Reactor

Calvert Cliffs No. 1

Fig. 4. Filter Data node with the Reactor module selected. The desired reactor can be chosen via the bottom dropdown menu. Note the two output sockets.

Specific modules for group inclusion/exclusion include:

- **PCA Model:** This module uses modeling/training data to compute the PCA model parameters and stores the results in tables including PCA Model Overview, PCA Component Information, PCA Statistics, PCA Eigenvectors, and PCA Covariance. Also included is a table listing the modeling results used as input for each PCA model. The PCA model is calculated using MATLAB code that has been integrated into the Augment Data node.
- **Assign KNN Class:** This module creates a table containing unique class identifiers and a table of fuel pins assigned to each class.
- **KNN Training:** This module uses class fuel pin test results to create a series of KNN Training results tables, including KNN Training Statistics, KNN Included Classes, KNN Neighbor Information, KNN Success Information, KNN Training Class Statistics, KNN Goodness Values and KNN Goodness Thresholds. KNN training is performed using MATLAB code that has been integrated into the Augment Data node.

4.3. Filter data node

Data filters are placed within the network to split data into two separate flow paths based on the selection of filtering criteria. As such, a filter node, shown in Fig. 4, has one input socket and two output sockets. Filtering is accomplished with the creation of two filter tables. The first lists the database records that pass the criteria and the second lists those that do not. Although the result of the filter operations is two additional tables in the internal database, only one of the two is referenced on each of the outgoing sockets. DAE maintains the relational information necessary to apply the

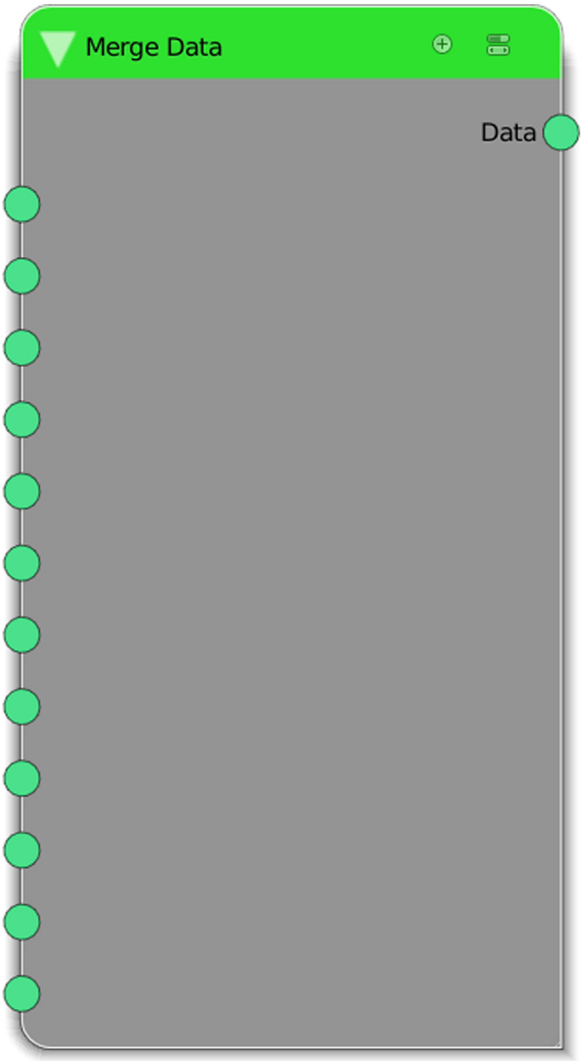


Fig. 5. Merge Data Node. Database references on each of the twelve input sockets are made available on the output socket.

correct filter information (passed or not passed) downstream of the filter node.

The common PCA/KNN Filter modules are:

- **Reactor:** This module creates filter tables based on the selected reactor.
- **Test:** This module creates filter tables based on the selected test or measurement.
- **Fuel Pin:** This module creates filter tables based on selected fuel pin test result criteria.

The PCA-specific module is:

- **PCA Retained Factors:** This module filters the PCA Model results based on the number of retained factors.

The KNN-specific module is:

- *k* **Nearest Neighbors:** This module filters the KNN Training results based on the number of nearest neighbors.

4.4. Merge data node

Merging of data streams within the network is accomplished via a Merge Data node, shown in Fig. 5. No node selections are present (or necessary) on this node type. It has twelve input sockets and one output socket. Functionally, the node acts to make all input database references on the input sockets known on the output socket. By merging data in this way, a logical operation of combining the results of parallel filter paths is possible. In addition, this node can be used to merge two initially independent data source streams into one for downstream processing of both sets of data. In particular, this merging allows access to multiple input sets for creating relationships between them.

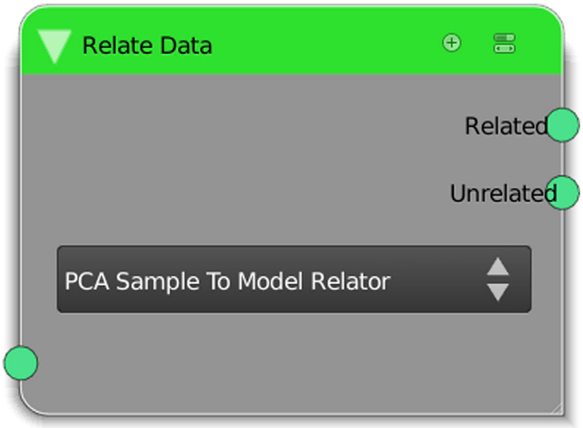


Fig. 6. Relate Data node with the PCA Sample to Model module selected in the dropdown menu. Note the two output sockets.

4.5. Relate data node

Data relater nodes, as shown in Fig. 6, are used to create relationships between initially independent sets of source data (or their derived augmentation data). These nodes have a single input socket and two output sockets. Prior to processing, an upstream merge node is required to direct the independent data sets to a single input socket. One output socket provides information on the relationships formed by the node. The other socket provides information regarding the unrelated input data sets. For the socket outputting relation information, a new table is created with database columns relating information from both sets. When information from an input set is not included in the new relationship table, corresponding entries are made in new "filter-like" tables listing the non-related source records. Two such filter-like tables are created and are used to filter the input sets to only those not included on the relation output socket.

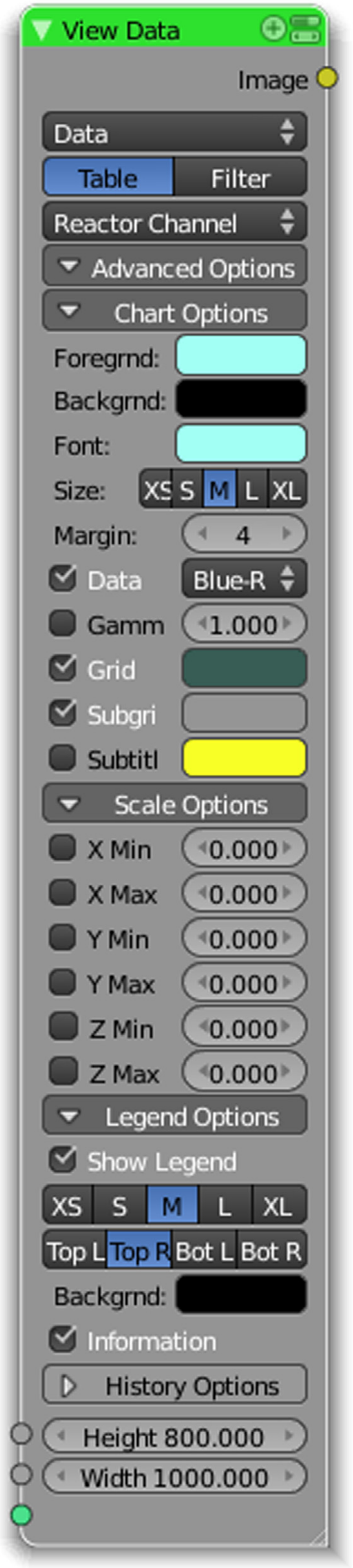
Each output socket contains only the references to the newly created tables that are relevant to each output. The "related" socket will contain a reference to the new relation table while the "not related" socket will reference the two newly created filter tables. This allows the user to easily work with the new relationships or identify information from the initial data sets from which no relationship was established.

The PCA-specific module is:

- **PCA Sample to Model:** This module computes a series of metrics that represent the likelihood that each questioned/test sample is represented by each PCA model. These results are included in a relationship table. These calculations are performed using MATLAB code that has been integrated into the Relate Data node.

The KNN-specific module is:

- **KNN Sample to Training:** This module computes a series of metrics that represent the likelihood that each questioned/test sample is represented by each KNN Class. These results are included in a relationship table. These calculations are performed using MATLAB code that has been integrated into the Relate Data node.



View Data

Image

Data

Table | Filter

Reactor Channel

Advanced Options

Chart Options

Foregrnd:

Backgrnd:

Font:

Size: XS S M L XL

Margin: 4

☑ Data Blue-R

□ Gamm 1.000

☑ Grid

☑ Subgri

□ Subtitl

Scale Options

□ X Min 0.000

□ X Max 0.000

□ Y Min 0.000

□ Y Max 0.000

□ Z Min 0.000

□ Z Max 0.000

Legend Options

☑ Show Legend

XS S M L XL

Top L Top R Bot L Bot R

Backgrnd:

☑ Information

History Options

Height 800.000

Width 1000.000

Fig. 7. View Data Node (fully expanded). Note that the image height and width can be defined using two input sockets.

4.6. *View data node*

The View Data node (shown fully expanded in Fig. 7) is used to host algorithms designed to create visual representations of the network data. It has a single input socket and a single output socket. This node architecture is unique, however, in that the output socket type is an image. Images are one of the three basic Blender outputs (i.e., value, vector, and image). Once an image is created and output by a View Data node, the image can be directed to any of the existing Blender nodes that accept incoming images. This allows for post-processing of the created images (such as overlapping or differencing).

The common PCA/KNN modules are:

- **Data:** This module creates tabular information from internal database tables for inspection.
- **Fuel Pin Test:** This module creates an x–y plot of one variable versus another variable for a designated set of training data. The training data are labeled according to class. This plot is generated using IDL code that has been integrated in the View Data node.

The PCA-specific modules are:

- **PCA Sample to Model:** This module creates a tabular representation of comparison results of questioned/test samples to PCA models.
- **PCA Correlation Metrics:** This module creates a plot of the PCA correlation metrics.
- **PCA Eigenvalues:** This module creates a plot of the PCA eigenvalues. This plot is generated using IDL code that has been integrated into the View Data node.
- **PCA Loadings:** This module creates a plot of the PCA loadings. This plot is generated using IDL code that has been integrated into the View Data node.

The KNN-specific modules are:

- **KNN Sample to Training:** This module creates a tabular representation of the comparison results of questioned/test samples to KNN training classes.
- **KNN Success:** This module creates a plot of KNN success, defined as the percentage of training samples that are correctly classified as a function of the number of nearest neighbors. This plot is generated using IDL code that has been integrated into the View Data node.

4.7. *Compose view node*

The Compose View Node is used to create a mosaic image composed of multiple incoming images. The node provides for horizontal or vertical stacking of images. Nesting of this node type can be used to create complex layouts from many images. Compose View nodes have up to six input sockets and one output. All sockets are of image type.

4.8. *Date/time node*

The Date/Time node is used in conjunction with source nodes to import specific ranges of time-stamped data. This data type requires a new type of socket information to transfer date and time information from node to node. This DAE node type was not used in the analysis of SFCOMPO.

4.9. *Example DAE PCA analysis*

Fig. 8 depicts an example PCA analysis of the SFCOMPO and simulated SNF data performed using DAE. On the far left, the Model Data Source module is used to load the SFCOMPO model data, and the Sample Data Source module is used to load the simulated SNF questioned/test data. A Merge Data node is then used to merge these two data streams. Five Filter Data nodes are next employed to limit the set of analyzed variables to the five isotopic ratios: ^{235}U/^{238}U, ^{236}U/^{238}U, ^{240}Pu/^{239}Pu, ^{241}Pu/^{239}Pu, and ^{242}Pu/^{239}Pu. As

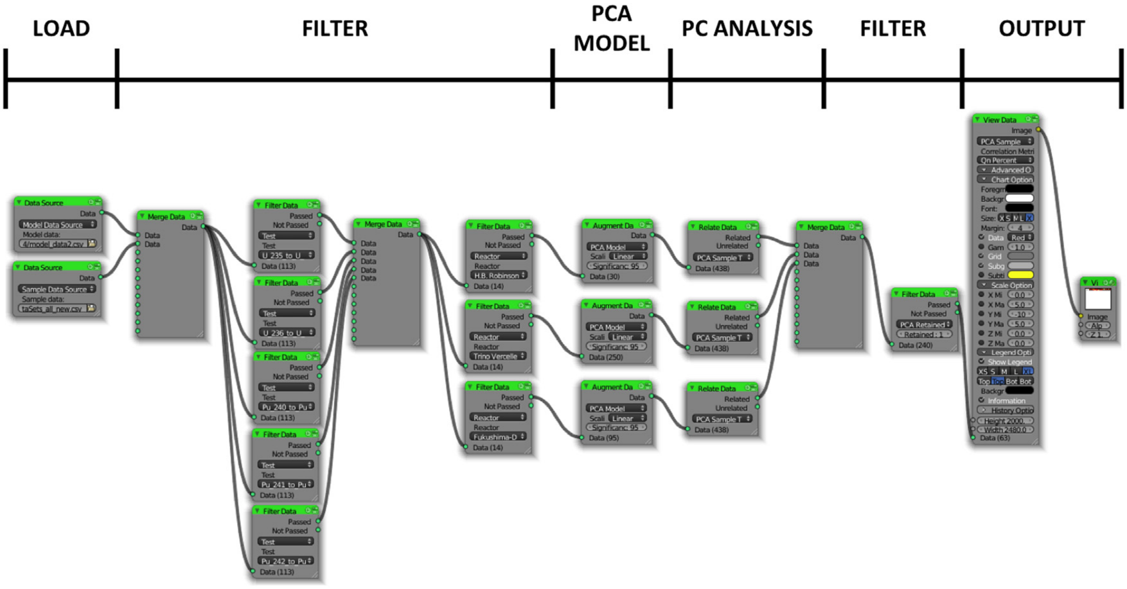


Fig. 8. Example PCA analysis of SFCOMPO and simulated SNF data performed using DAE.

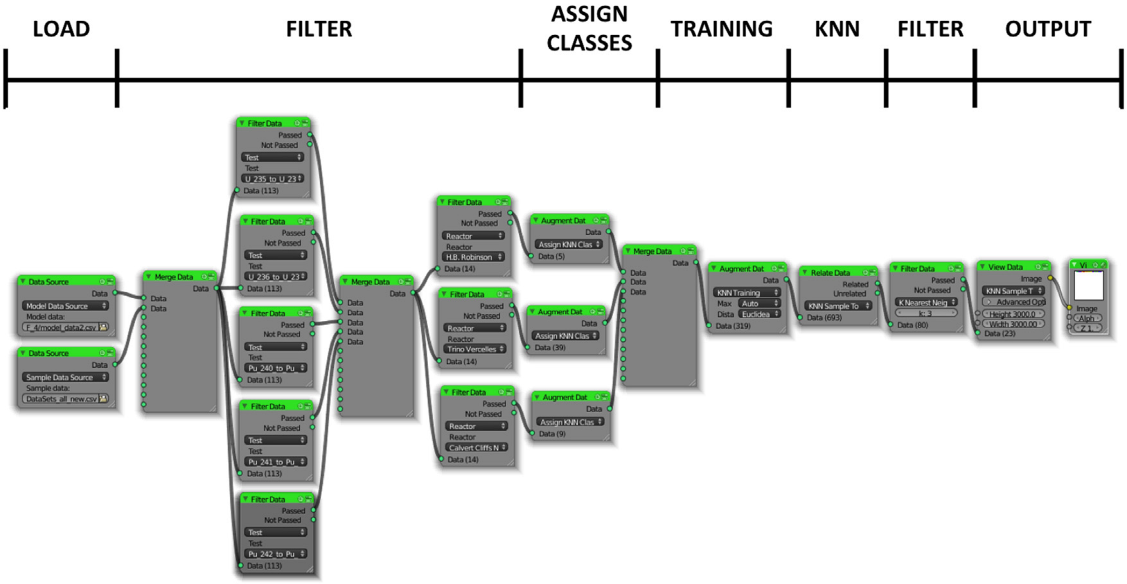


Fig. 9. Example KNN analysis of SFCOMPO and simulated SNF data performed using DAE.

a fourth step in the analysis chain, a Merge Data node is used to merge these five isotopic ratios. Fifth, three Filter Data/Reactor modules are used to limit the data to the selected H.B. Robinson Unit 2, Trino Vercellese, and Fukushima-Daini-2 reactors. Sixth, three Augment Data/PCA Model modules are utilized to construct PCA models for the selected H.B. Robinson Unit 2, Trino Vercellese, and Fukushima-Daini-2 reactors. Seventh, three Relate Data/PCA Sample to Model modules are used to compare the simulated SNF test data to the three constructed PCA models. Eighth, a Merge Data node is employed to merge these PCA model results for the SNF test data. Ninth, a Filter Data/PCA Retained Factors module is utilized to filter the PCA model results to one retained PC. Tenth, a View Data/PCA Sample to Model module is used to create a tabular representation of the comparison results of the simulated SNF test samples to the three PCA models. On the far right, a Compose View node is used to generate a mosaic image of the results.

4.10. *Example DAE KNN analysis*

Fig. 9 depicts an example KNN analysis of the SFCOMPO and simulated SNF data performed using DAE. The first five steps in the analysis chain match those employed in PCA. In the sixth step, three Augment Data/Assign KNN Class modules are used to assign unique identifiers to the three reactor classes. Seventh, a Merge Data node is employed to merge the data from the three reactor classes. Eighth, an Augment Data/KNN Training module is used to create a series of KNN Training results tables employing the SFCOMPO model data. Ninth, a Relate Data/KNN Sample to Training module is utilized to compute a series of metrics representing the likelihood that each simulated SNF questioned/test sample is a member of each of the three KNN reactor classes. Tenth, a Filter Data/*k* Nearest Neighbors module is used to filter the KNN training results to three nearest neighbors. Eleventh, a View Data/KNN Sample to Training module is employed to create a

Table 2
Q statistic-based probabilities calculated for simulated samples 1–20 for each of the 14 autoscaled PCA models. Entries greater than or equal to the exclusion threshold of 0.01 are underlined.

Sample/RF	A	A & G	B	C	D	E & I	F	G	H	J	K	L	M	N
	2	2	1	1	1	1	1	1	1	2	1	2	1	3
1	0.10	0.58	0.00	0.00	0.64	0.00	0.00	0.43	0.00	0.00	0.00	0.00	0.01	0.00
2	0.08	0.31	0.00	0.00	0.40	0.00	0.00	0.61	0.00	0.00	0.01	0.00	0.00	0.00
3	0.15	0.45	0.00	0.00	0.20	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00
4	0.05	0.42	0.00	0.00	0.55	0.00	0.00	0.45	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.57	0.01	0.00	0.00	0.00	0.00	0.03	0.02	0.43	0.00
6	0.00	0.00	0.00	0.00	0.10	0.51	0.00	0.00	0.00	0.00	0.02	0.03	0.75	0.00
7	0.00	0.00	0.00	0.00	0.44	0.01	0.00	0.00	0.00	0.00	0.10	0.04	0.30	0.00
8	0.00	0.00	0.00	0.00	0.25	0.09	0.00	0.00	0.00	0.00	0.10	0.06	0.46	0.00
9	0.00	0.00	0.00	0.00	0.04	0.28	0.00	0.00	0.00	0.00	0.00	0.03	0.64	0.00
10	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00
11	0.00	0.00	0.00	0.00	0.20	0.11	0.00	0.00	0.00	0.00	0.07	0.07	0.31	0.00
12	0.00	0.00	0.00	0.00	0.29	0.03	0.00	0.00	0.00	0.00	0.08	0.06	0.16	0.00
13	0.00	0.00	0.00	0.00	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.84	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
16	0.00	0.00	0.00	0.00	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.17	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00
18	0.00	0.00	0.00	0.00	0.76	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	0.00	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00
20	0.00	0.00	0.00	0.00	0.66	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

tabular representation of the comparison results of the simulated SNF questioned/test samples to the three KNN training classes. On the far right, a Compose View node is used to generate a mosaic image of the results.

5. Results and discussion

5.1. PCA results

Tables 2–4 present the *Q* statistic, Hotelling's T^2 statistic, and Hawkins' T_H^2 statistic-based probability values, respectively, for simulated questioned/test samples 1 through 20 calculated in DAE for each of the 14 autoscaled PCA models. Autoscaling refers to the process in which a data matrix is transformed through centering, where the mean of each variable is subtracted from all of its elements, followed by variance scaling, achieved by dividing each element by the standard deviation of that variable. In the tables, a probability value greater than or equal to the exclusion threshold of 0.01 indicates that the simulated sample is considered to be consistent with the modeled group, making a case for group inclusion. These tables are representative of outputs provided by DAE through the View Data/PCA Sample to Model module.

With regard to the *Q* statistic-based probabilities (Table 2), the following group inclusion/exclusion results are achieved:

- Simulated samples 1 through 4 exceed the exclusion threshold of 0.01 for (1) the Calvert Cliffs No. 1 model, (2) the combined Calvert Cliffs No. 1/H.B. Robinson Unit 2 model, (3) Fukushima-Daini-2 model, and (4) H.B. Robinson Unit 2 model. The true reactor of origination for simulated samples 1 through 4 is H.B. Robinson Unit 2.
- All or a subset of simulated samples 5 through 12 exceed the exclusion threshold for (1) the Fukushima-Daini-2 model, (2) the Genkai-1/Mihama-3 model (6 of 8 samples), (3) the Obrigheim model (6 of 8 samples), (4) the Takahama-3 model (7 of 8 samples), and (5) the Trino Vercellese model. The true reactor of origination for simulated samples 5 through 12 is Trino Vercellese.
- All or a subset of simulated samples 13 through 20 exceed the exclusion threshold for (1) the Fukushima-Daini-2 model, (2) the Takahama-3 model (1 of 8 samples), and (3) the Trino Vercellese model (3 of 8 samples). The true reactor of origination for simulated samples 13 through 20 is Fukushima-Daini-2.

Thus, for all 20 simulated samples the *Q* statistic correctly excludes a large percentage of reactor groups and correctly includes the true reactor of origination.

With regard to the Hotelling's T^2 statistic-based probabilities (Table 3), the following group inclusion/exclusion results are achieved:

- All or a subset of simulated samples 1 through 4 exceed the exclusion threshold for 12 of the 14 PCA reactor models.
- All or a subset of simulated samples 5 through 12 exceed the exclusion threshold for 13 of the 14 PCA reactor models.
- All or a subset of simulated samples 13 through 20 exceed the exclusion threshold for 13 of the 14 PCA reactor models.

Accordingly, the group inclusion/exclusion results achieved by Hotelling's T^2 statistic are much more ambiguous. For all 20 simulated samples, Hotelling's T^2 statistic correctly includes the true reactor of origination, but excludes very few of the other reactor groups.

Finally, employing the Hawkins' T_H^2 statistic-based probabilities (Table 4), the following group inclusion/exclusion results are achieved:

- Simulated samples 1 through 4 exceed the exclusion threshold for (1) the Calvert Cliffs No. 1/H.B. Robinson Unit 2 model, and (2) the Fukushima-Daini-2 model.
- All or a subset of simulated samples 5 through 12 exceed the exclusion threshold for (1) the Calvert Cliffs No. 1/H.B. Robinson Unit 2 model (1 of 8 samples), (2) the Fukushima-Daini-2 model (6 of 8 samples), (3) the Genkai-1/Mihama-3 model (5 of 8 samples), (4) the Takahama-3 model (6 of 8 samples), and (5) the Trino Vercellese model (1 of 8 samples).

Table 3
Hotelling's T^2 statistic-based probabilities calculated for simulated samples 1–20 for each of the 14 autoscaled PCA models. Entries greater than or equal to the exclusion threshold of 0.01 are underlined.

Sample/RF	A	A & G	B	C	D	E & I	F	G	H	J	K	L	M	N
	2	2	1	1	1	1	1	1	1	2	1	2	1	3
1	0.09	0.14	0.25	0.02	0.24	0.58	0.14	0.18	0.00	0.00	0.08	0.25	0.85	0.09
2	0.71	0.91	1.00	0.09	0.88	0.66	0.37	0.85	0.00	0.00	0.92	0.20	0.03	0.00
3	0.88	0.90	0.60	0.22	0.70	0.37	0.07	0.38	0.00	0.00	0.53	0.10	0.00	0.00
4	0.35	0.55	0.60	0.05	0.54	0.99	0.88	0.60	0.00	0.00	0.42	0.24	0.20	0.05
5	0.23	0.07	0.86	0.03	0.66	0.87	0.68	0.86	0.00	0.00	0.66	0.85	0.14	0.00
6	0.05	0.05	0.10	0.01	0.10	0.28	0.01	0.05	0.00	0.00	0.01	0.31	0.40	0.00
7	0.29	0.15	0.49	0.02	0.42	0.81	0.54	0.38	0.00	0.00	0.25	0.74	0.46	0.51
8	0.13	0.13	0.21	0.01	0.20	0.47	0.07	0.12	0.00	0.00	0.05	0.49	0.88	0.00
9	0.01	0.01	0.05	0.00	0.06	0.17	0.00	0.03	0.12	0.00	0.00	0.19	0.17	0.00
10	0.01	0.00	0.04	0.00	0.04	0.12	0.00	0.02	0.87	0.00	0.00	0.12	0.09	0.00
11	0.09	0.10	0.16	0.01	0.16	0.39	0.04	0.09	0.00	0.00	0.03	0.42	0.68	0.00
12	0.13	0.15	0.20	0.01	0.20	0.47	0.07	0.12	0.00	0.00	0.05	0.45	0.85	0.00
13	0.01	0.00	0.12	0.30	0.17	0.09	0.00	0.05	0.00	0.12	0.03	0.16	0.00	0.00
14	0.02	0.00	0.25	0.13	0.38	0.19	0.01	0.11	0.00	0.01	0.12	0.39	0.00	0.00
15	0.13	0.05	0.28	0.01	0.28	0.59	0.16	0.19	0.00	0.00	0.10	0.67	0.82	0.00
16	0.08	0.02	0.79	0.04	0.94	0.56	0.19	0.60	0.00	0.00	0.81	0.89	0.02	0.01
17	0.14	0.06	0.33	0.01	0.33	0.66	0.24	0.24	0.00	0.00	0.13	0.74	0.68	0.01
18	0.01	0.00	0.29	0.08	0.50	0.25	0.01	0.15	0.00	0.01	0.19	0.57	0.00	0.00
19	0.03	0.00	0.82	0.01	0.66	0.91	0.73	0.85	0.00	0.00	0.63	0.88	0.16	0.04
20	0.09	0.02	0.73	0.05	0.88	0.52	0.15	0.53	0.00	0.00	0.72	0.84	0.01	0.00

Table 4
Hawkins' T_H^2 statistic-based probabilities calculated for simulated samples 1–20 for each of the 14 autoscaled PCA models. Entries greater than or equal to the exclusion threshold of 0.01 are underlined.

Sample/RF	A	A & G	B	C	D	E & I	F	G	H	J	K	L	M	N
	2	2	1	1	1	1	1	1	1	2	1	2	1	3
1	0.00	0.67	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.18	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.61	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.60	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.01	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
6	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.03	0.17	0.00
7	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00
8	0.00	0.00	0.00	0.00	0.09	0.05	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
11	0.00	0.00	0.00	0.00	0.12	0.07	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00
12	0.00	0.00	0.00	0.00	0.17	0.02	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00
13	0.00	0.00	0.00	0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
14	0.00	0.00	0.00	0.00	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	0.00	0.00	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.01	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
17	0.00	0.00	0.00	0.00	0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
19	0.00	0.00	0.00	0.00	0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
20	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

- All or a subset of simulated samples 13 through 20 exceed the exclusion threshold for (1) the Fukushima-Daiichi-3 model (1 of 8 samples), and (2) the Fukushima-Daini-2 model.

The group inclusion/exclusion results obtained by Hawkins' T_H^2 statistic are inferior to those achieved by the Q statistic. Hawkins' T_H^2 statistic includes the true reactor of origination for only 13 of the 20 simulated samples, but does successfully exclude a large percentage of the reactor groups.

One of the challenges encountered in constructing the PCA models was the limited number of samples available for some of the reactors. As a case in point, only 6 samples are available for the Cooper reactor, resulting in a relatively high level of uncertainty in the calculated Cooper PCA model and associated control limits for the Q, Hotelling's T^2, and Hawkins' T_H^2 statistics. Additionally, fewer than 10 samples are available for the Calvert Cliffs No. 1, H.B. Robinson Unit 2, Genkai-1, and Mihama-3 reactors. In constructing the PCA models, samples from related sets of reactors, such as Genkai-1 and Mihama-3, were combined in an attempt to reduce model uncertainty. However, superior group inclusion/exclusion results would likely have been achieved if additional samples were available to model the reactors.

A notable trend observed in the inspection of Tables 2–4 is the consistent inclusion of the majority or all of the 20 simulated samples in the Fukushima-Daini-2 model. This trend is attributable to the presence of three outliers among the set of Fukushima-Daini-2 samples used for modeling (see Fig. 1). These outliers effectively expand the Fukushima-Daini-2 PCA model, resulting in the inclusion of many or all of the simulated samples in the Fukushima-Daini-2 group. In order to avoid such issues, the ability to remove outliers from a training set prior to model construction is planned for an updated version of the DAE package.

In practice, it would be preferable to utilize a single index in performing PCA group inclusion/exclusion for a questioned sample as opposed to monitoring multiple indices such as the Q statistic, Hotelling's T^2 statistic, and Hawkins' T_H^2 statistic. In the context of fault detection in the monitoring of industrial processes, Yue and Qin proposed a combined index that integrates information contained in both the Q statistic and Hotelling's T^2 statistic [21]. Their motivation for combining the Q statistic and Hotelling's T^2 statistic into a single index is that these two indices provide complementary information—for a questioned sample, the Q statistic detects unusual variability outside the PCA model space, while Hotelling's T^2 statistic detects unusual variability within the PCA model space. A combined index that integrates the Q statistic and Hotelling's T^2 statistic is planned for an updated version of the DAE package.

5.2. KNN results

In performing KNN, the training set was constructed from samples from the following individual or combined sets of related reactors: (1) Calvert Cliffs No. 1/H.B. Robinson Unit 2, (2) Cooper, (3) Fukushima-Daiichi-3, (4) Fukushima-Daini-2, (5) Genkai-1/Mihama-3, (6) Gundremmingen, (7) JPDR, (8) Monticello, (9) Obrigheim, (10) Takahama-3, (11) Trino Vercellese, and (12) Tsuruga-1. These 12 individual or related sets of reactors formed the groups used for KNN classification. The set of 5 isotopic ratios used in KNN classification, namely ^{235}U/^{238}U, ^{236}U/^{238}U, ^{240}Pu/^{239}Pu, ^{241}Pu/^{239}Pu, and ^{242}Pu/^{239}Pu, was autoscaled. The k value of 1 employed in the KNN classification of the 20 simulated samples was selected by leave-one-out cross validation of the training set, where it was found that a value of $k = 1$ maximized the classification accuracy for the training set [15]. KNN group inclusion/exclusion of the 20 simulated samples was performed utilizing the goodness value criterion. Specifically, a simulated sample was excluded from membership in its predicted group if its goodness value exceeded the maximum goodness value for the training samples that are members of that group.

Fig. 10 summarizes the group inclusion/exclusion results obtained by KNN for the 20 simulated samples. In Fig. 10, a goodness value is listed for a given simulated sample and reactor group in the event that the simulated sample was classified by KNN as a member of that group, and the goodness value for the simulated sample was less than or equal to the maximum goodness value for training samples that are members of that group. An empty cell for a given simulated sample and reactor group indicates that either the simulated sample was not classified by KNN as a member of that group, or the simulated sample was excluded from membership in its predicted group as its goodness value exceeded the maximum goodness value for training samples that are members of that group. An inspection of Fig. 10 reveals the following KNN group inclusion/exclusion results for the 20 simulated samples:

- Simulated samples 1 through 4 are classified as members of the Calvert Cliffs No. 1/H.B. Robinson Unit 2 group. These classifications are consistent with the true reactor of origination for simulated samples 1 through 4 of H.B. Robinson Unit 2.
- Simulated samples 5 through 9 and 12 are classified as members of the Trino Vercellese group, while simulated samples 10 and 11 are classified as members of the Genkai-1/Mihama-3 group. The true reactor of origination for simulated samples 5 through 12 is Trino Vercellese.
- Simulated samples 13, 14, 18, and 19 are classified as members of the Fukushima-Daini-2 group, while simulated samples 15 through 17 and 20 are excluded from all of the known groups based on the goodness value criterion. The true reactor of origination for simulated samples 13 through 20 is Fukushima-Daini-2.

In summary, 14 of the 20 simulated samples are correctly classified by KNN to their true reactor of origination, 2 of the 20 simulated samples are incorrectly classified to a reactor group, and 4 of the 20 simulated samples are excluded from all of the known reactor groups. The KNN group inclusion/exclusion results for the simulated samples are impressive considering the overlap between reactors in the variable space (see Fig. 1), the limited number of training samples available for some reactors, and the presence of outliers in the training set.

6. Conclusions and future work

The DAE software package is a powerful tool for systematic nuclear forensic analyses. The capability for interfacing via a variety of programming languages and the modular work flow provides a platform for future development of advanced group inclusion/exclusion methods.

The design of DAE yields an extensible software framework that allows for the straightforward integration of new scientific modules. The integration of these new modules involves a basic four-step process that includes: (1) the schema design of the database tables required to store all output information, (2) the specification of existing database information required as input to the new modules, (3) the design of the visual module layout that presents the control parameters to the user for manipulation, and (4) the incorporation of new mathematical routines required for the transformation of input data into output data. The new output data is stored within the internal database for use by other downstream DAE network modules.

DAE modules do not interact directly with each other; rather they interact indirectly via access to the common database of stored information. This architectural feature allows for parallel module development efforts by various contributors without the need for constant coordination. The database schema itself provides all necessary interface information. New modules can selectively use only the existing database information required for operation, and each developer can independently design the table schemas for output from their new module. Once the new table schemata are finalized and shared, other developers can design modules to utilize the newly created data.

Several enhancements are planned for the DAE software package, including:

- The incorporation of additional data preprocessing methods such as range scaling [22]. Data preprocessing is employed to transform the raw data (e.g., trace elements and isotopics) to new units or scales to improve group inclusion/exclusion performance.

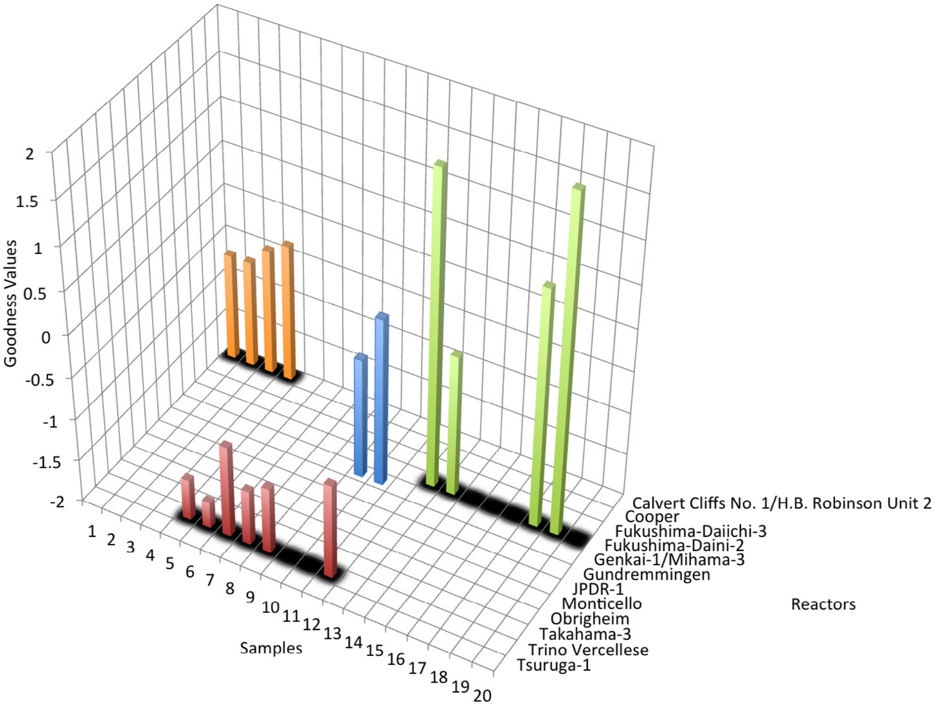


Fig. 10. Group inclusion/exclusion results obtained by KNN with goodness value criterion for 20 simulated samples.

- The integration of additional data visualization capabilities. Data visualization involves plotting the raw or preprocessed data to identify differences between groups of nuclear materials and the similarity of questioned materials to known groups.
- The ability to remove outlier samples from a training set prior to group inclusion/exclusion model construction.
- The incorporation of a combined index that integrates the Q statistic and Hotelling's T^2 statistic for PCA-based group inclusion/exclusion.
- The incorporation of additional group inclusion/exclusion methods such as one-class support vector machines (SVM) [23]. In selecting new group inclusion/exclusion methods for incorporation in the DAE software package, emphasis will be placed on methods that provide an estimate on the degree of certainty of classification for a questioned material.
- The incorporation of group inclusion/exclusion performance evaluation metrics, such as accuracy and precision, which will enable the quantitative comparison of methods.

Acknowledgments

The authors acknowledge the late Richard McKnight at Argonne National Laboratory for generating the simulated samples used in evaluating the group inclusion/exclusion methods.

The authors thank the U.S. Department of Homeland Security/National Technical Nuclear Forensics Center for funding and supporting this work.

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.

This material is based upon work supported in part by the Department of Energy National Nuclear Security Administration through the Nuclear Science and Security Consortium under Award Number DE-NA0000979.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

[1] K. Moody, P. Grant, I. Hutcheon, Nuclear Forensic Analysis, CRC Press, 2005. URL https://books.google.com/books?id=Q9mgDnWoPLYC.
[2] K. Mayer, M. Wallenius, I. Ray, Analyst 130 (2005) 433–441. http://dx.doi.org/10.1039/B412922A.
[3] J. Borgardt, F. Wong, J. Nucl. Mater. Manage. XLII (2014) 4–11.
[4] G. Griffiths, E. Loi, D. Boardman, D. Hill, K.L. Smith, J. Nucl. Mater. Manage. XLII (4) (2014) 12–23.
[5] J.E.S. Sarkis, I.C.A.C. Bordon, R.C.B. Pestana, R.C. Marin, J. Nucl. Mater. Manage. XLII (4) (2014) 24–30.
[6] A. El-Jaby, R. Kosierb, F. Doucet, I. Dimayuga, G. Edwards, D. Barber, S. Corbett, D. Wojtaszek, J. Nucl. Mater. Manage. XLII (4) (2014) 31–39.
[7] Y. Kimura, N. Shinohara, Y. Funatake, J. Nucl. Mater. Manage. XLII (4) (2014) 40–45.
[8] A.N. Nelwamondo, A.M. Bopape, J.H. Bohlolo, K.K. Nkuna, J. Nucl. Mater. Manage. XLII (4) (2014) 46–54.
[9] A.J. Heydon, C.A. Cooper, P. Thompson, P.G. Turner, R. Gregg, K.W. Hesketh, A.E. Jones, J.Y. Goulermas, J. Nucl. Mater. Manage. XLII (4) (2014) 55–64.
[10] É. Kovács-Széles, S. Szabó, T.C. Nguyen, J. Nucl. Mater. Manage. XLII (4) (2014) 65–69.
[11] A. Axelsson, H. Ramebäck, B. Sandström, J. Nucl. Mater. Manage. XLII (4) (2014) 70–75.
[12] M. Wallenius, Z. Varga, K. Mayer, J. Nucl. Mater. Manage. XLII (4) (2014) 76–82.
[13] SFCOMPO - Spent fuel isotopic composition database, 2015. URL http://www.oecd-nea.org/sfcompo/.
[14] J.E. Jackson, A User's Guide to Principal Components, Wiley-Interscience, New York, NY, 2003.
[15] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, John Wiley & Sons, 1998.
[16] S.J. Qin, J. Chemometr. 17 (2003) 480–502.
[17] D.M. Hawkins, J. Amer. Statist. Assoc. 69 (346) (1974) 340–344.
[18] B.V. Dasarathy, IEEE Trans. Pattern Anal. Mach. Intell. 2 (1980) 67–71.
[19] Blender Online Community, Blender - a 3D modeling and rendering package, Blender Foundation, Blender Institute, Amsterdam, 2015. URL http://www.blender.org.
[20] Python Software Foundation, Wilmington, DE, sqlite3 — DB-API 2.0 interface for SQLite databases, 2016. URL https://docs.python.org/2/library/sqlite3.html.
[21] H.H. Yue, S.J. Qin, Ind. Eng. Chem. Res. 17 (8–9) (2003) 480–502.
[22] M.A. Sharaf, D.L. Illman, B.R. Kowalski, Chemometrics, Wiley-Interscience, 1986.
[23] S.S. Khan, M.G. Madden, Knowl. Eng. Rev. 29 (3) (2014) 345–374.