Scenario Aggregation and Analysis via Mean-Shift Methodology

Diego Mandelli[a*], Alper Yilmaz[b], Tunc Aldemir[a], Richard Denning[a]

[a] The Ohio State University, Nuclear Eng. Program; Columbus (OH), USA
[b] The Ohio State University, Civil, Environmental Eng. and Geodetic Sciences; Columbus (OH), USA

Abstract: A challenging aspect of dynamic methodologies, such as the Dynamic Event Tree (DET) methodology, is the large number of scenarios generated for a single initiating event. Such large amounts of information can be difficult to organize in order to extract useful information. The scenario dataset is composed of scenarios which contain information on the system components and the system process variables, such as values of pressures and temperatures for the reactor coolant system and the containment throughout the time period of the transient. In order to facilitate analysis, it can be fruitful to accomplish two tasks: i) identify the scenarios that have a "similar" behavior (i.e. identify the most evident classes), and, ii) decide, for each event sequence, to which class it belongs (i.e., classification). It is shown how it is possible to accomplish these two tasks using the Mean-Shift Methodology. The Mean-Shift methodology is a kernel-based, non-parametric density estimation technique that is used to find the modes of an unknown distribution, which corresponds to regions with highest data density. The methodology is illustrated by applying it to the DET analysis of a simple level controller.

Keywords: Scenario Classification, Dynamic PRA, Pattern Recognition, Data Analysis.

1. INTRODUCTION

The Event Tree (ET)/Fault Tree (FT) approach is a traditional tool for probabilistic safety/risk assessment (PSA/PRA) not only for nuclear systems but also for the aerospace, chemical and transportation industries. However, several concerns have been raised about the capability of the ET/FT approach to treat the coupling between the plant physical processes and triggered or stochastic logical events [1] which can have significant impact on the consequences of upset conditions and their frequencies. Another concern is the contribution of epistemic uncertainties to the ordering of events and consequences of upset conditions. As discussed in [2], a safety methodology has to be able to:

- model the dynamics of the system and, hence, needs to be coupled with system or plant simulators,
- model the exact time scale of the accident,
- model the change of hardware component states,
- model human interaction with the system dynamics, and,
- handle epistemic and aleatory uncertainties.

Dynamic PSA/PRA methodologies respond to these needs by using advanced system simulators to identify the timing of events and to account for the coupling between triggered and/or stochastic events [3, 4].

A challenging aspect of dynamic methodologies, such as the DET methodology [3], is the large number of scenarios generated for a single initiating event. Such large amounts of information can be difficult to organize for tractable analysis. In particular, as part of the PSA/PRA framework, it is important to identify the main scenario evolutions and the main risk contributors for each initiating event. In this work, we want to address this problem of data analysis by aggregating the scenarios into classes (or clusters) and analyzing the properties of the scenarios of each cluster.

[*] Reference Author: Diego Mandelli; e-mail: mandelli.1@osu.edu
Nuclear Engineering Program, The Ohio State University 201 W. 19th Ave., 43210 Columbus (OH), USA.

By scenario clustering we mean two actions:

1. Identify the scenarios that have a "similar" behavior (i.e. identify the most evident classes)
2. Decide for each event sequence which class it belongs (i.e., classification)

When dealing with nuclear transients, it is possible to analyze the set of scenarios in two possible modes:

1. *End State Analysis*: Classify the scenarios into clusters based on the end state of the scenarios
2. *Transient Analysis*: Classify the scenarios into clusters based on the time evolution of the scenarios

While the first mode has been widely used in the ET/FT analysis, the second one is only starting to be considered in the recent years [5].

In this paper we will present a methodology that falls in the second category. In particular, we will show how it is possible to classify transients generated by a DET algorithm [6].

2. THE MEAN-SHIFT METHODOLOGY (MSM)

The methodology that is presented here is based on the Mean-Shift algorithm which has been described first in [7]. The MSM is a non parametric iterative procedure that shifts each data point to the average of data points in its neighborhood in order to determine the cluster centers and to assign each point to one cluster center only. By cluster center we mean a region with high observation density (i.e., the modes of the data set).

In order to show how this methodology actually works, an example is illustrated here. Let us consider a system whose dynamic can be described through 3 variables: time (*t*), pressure (*p*) and temperature (*T*). Several scenarios generated as a consequence of a DET analysis are shown Figure 1 which illustrates how these scenarios is may be distributed in the state space.

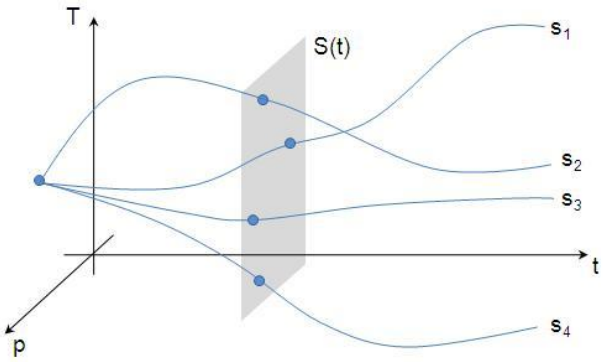


Figure 1: Representation of several scenarios generated by a DET in a 3-dimensional space.

Without loss of generality, let us consider only the state of these scenarios at a particular time instant. This operation is represented in Figure 1 as the projection of the scenarios onto the plane *S(t)*.

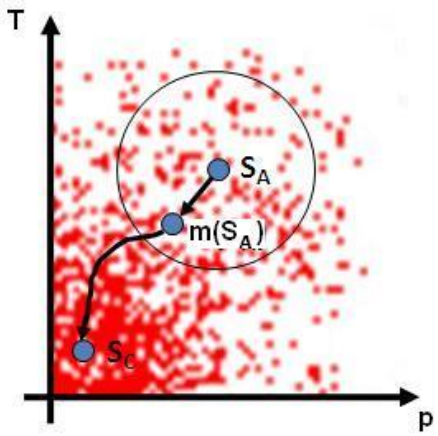


Figure 2: Determination of a cluster centre in a 2-dimensional space using a Mean-Shift algorithm.

At this point, the new data set consists of points distributed in a two dimensional space (i.e., $\mathbf{R}^2$). The idea behind the MSM is to determine the cluster centers and to assign each point to one cluster center only. By cluster center we intend a region with high point density. Starting form a generic point (i.e., point S_A in Figure 2), the algorithm associates a circle (or a sphere, depending on the number of dimensions of the state space) centered in that point. The radius of this area is identified as the bandwidth (*BW*). The idea is to consider all the points that are inside this circle and determine the center of mass of these points (point $m(s_A)$ in Figure 2). The center of mass is determined from

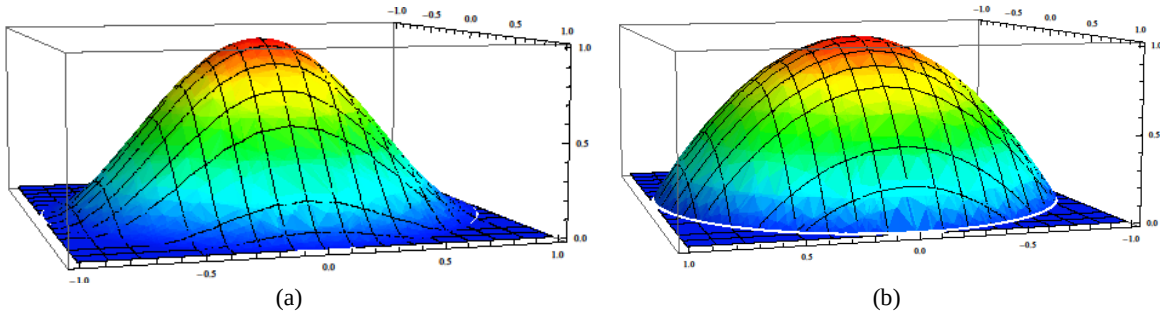
$$m(s_A) = \frac{\sum_{s \in S} K(s - s_A)s}{\sum_{s \in S} K(s - s_A)} \tag{1}$$

where the function *K(x)* is often referred to as the Kernel.

The purpose of *K(x)* is to assign different weights to different points during the estimation of the center of mass. Several Kernels can be used as illustrated in [9] and shown in Figure 3:

$$\text{Epanechnikov Kernel: } K(x) = \begin{cases} (1 - \|x\|^2) & \text{if } \|x\| \leq BW \\ 0 & \text{if } \|x\| > BW \end{cases} \tag{2}$$

$$\text{Bi-weighted Kernel: } K(x) = \begin{cases} (1 - \|x\|^2)^2 & \text{if } \|x\| \leq BW \\ 0 & \text{if } \|x\| > BW \end{cases} \tag{3}$$



(a) (b)

Figure 3: Graphical representation of the: (a) 2-D Bi-weighted and (b) the Epanechnikov Kernel

The algorithm then moves the original point S_A into the calculated position (point $m(s_A)$ in Figure 2) and repeats the calculation of the center of mass for the points included in the circle having identical value of bandwidth but now centered on $m(s_A)$. This operation stops when the distance between the

new center of mass and the old one is below a fixed threshold (point S_A in Figure 2). When this condition is reached:

- Point S_C is considered the center of a cluster
- The original point S_A is uniquely associated to the cluster centered by point S_C.

When this procedure is repeated for all the points in the data set it is possible to obtain:

- the center of all the clusters and the list of all the points that belong to that specific cluster, and,
- the cluster to which each point belongs (as mentioned, each point belongs to one cluster only).

3. SYSTEM UNDER CONSIDERATION

The first test for this methodology has been the analysis of the transients generated by the DET applied to a simple level controller described in [7] (see Figure 4).

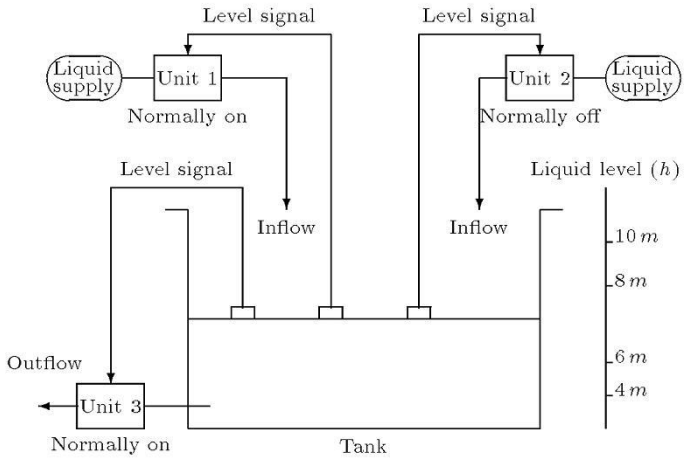


Figure 4: Scheme of the water heated tank.

The liquid level is actively controlled through the actuation of three components: two inlet pumps and one outlet valve, hereafter called Units 1, 2 and 3, respectively. Each unit is a multi-state component operating either correctly ON or OFF, or failed ON or failed OFF. At t=0, the system is assumed to be in its nominal state (Units 1, 2, 3 ON, OFF, ON, respectively), with equilibrium values of 30.93 ºC of the liquid temperature and 7 m of the level. The temperature of the liquid is assumed to have directl known effect on the failure rates of the components. A power source heats up the fluid to keep it at the nominal temperature. The control laws reported in [6] (see Table 1) act upon the state of the components to keep the liquid level between 4 and 10 m, the lower and upper safety thresholds, respectively.

Table 1: Control Laws for the level controller

	Control laws
1	If the liquid level *L* drops under 6 m, Units 1, 2, 3 are put respectively in state ON, ON and OFF (if they are not failed ON or OFF)
2	If the liquid level *L* rises above 8 m, Units 1, 2, 3 are put respectively in state OFF, OFF and ON (if they are not failed ON or OFF)

Thus, two possible Top Events need to be considered, i.e. dry-out (level < 4 m) and over-flow (level > 10 m). In this DET analysis, the branching is dictated by the failure of the three active components (i.e. Units 1, 2, and 3).

We then applied the Mean-Shift Methodology to the set of 619 transients generated by the DET.

4. SYSTEM ANALYSIS

Since our intent is to consider each transient as a single point in a multidimensional space, we converted each transient into a vector where level (L) and temperature (T) are sampled every hour. Thus, from a mathematical viewpoint, a generic transient $\boldsymbol{x}_i$ can be seen as following:

$$\boldsymbol{x}_i = [L(0), T(0), L(1), T(1), \cdots, L(n), T(n)] \tag{4}$$

where n represents the number of times the variables have been sampled.

When dealing with nuclear transients, the nature and the range of the variables of interest may differ significantly. This becomes particularly relevant when we define distance between points as specified in Eq. 2 and 3. The idea is to find the optimal representation of the points in a general n-dimensional space. In this respect, we decided to pre-process the data generated by the DET with a Principal Components Analysis (PCA) [10]. We implemented PCA by using the following procedure:

1. Consider the matrix the matrix $\boldsymbol{P}$ which contains all the data generated by the DET:

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_i \\ \vdots \\ \boldsymbol{x}_N \end{bmatrix} \tag{6}$$

2. Subtract the mean for each dimension
3. Determine the covariance matrix $\boldsymbol{C}$

$$\boldsymbol{C} = \frac{1}{N-1} \boldsymbol{P}\boldsymbol{P}^T \tag{5}$$

4. Evaluate the eigenvalues and eigenvectors of the covariance matrix $\boldsymbol{C}$
5. Project the original data $\boldsymbol{P}$ into the eigenvector space

We then performed the Mean-shift analysis of the data projected into the eigenvector space and converted the clusters generated obtained by the MSM into the original form.

5. RESULTS

Figures 5 and 6 show the cluster centers for the data generated by the DET for 2 different values of bandwidth *BW* (i.e., *BW* = 5, 2) for the 2 Top Events separately. As mentioned earlier, a cluster center can be viewed as the representative scenario for a subset of scenarios (i.e., a cluster of scenarios) where the size of the cluster itself depends on the chosen value of *BW*. With a broader value of *BW*, the algorithm identifies only 6 different clusters while a narrower value of *BW* is able to identify a larger numbers clusters (and hence refining the resolution of the clustering process).

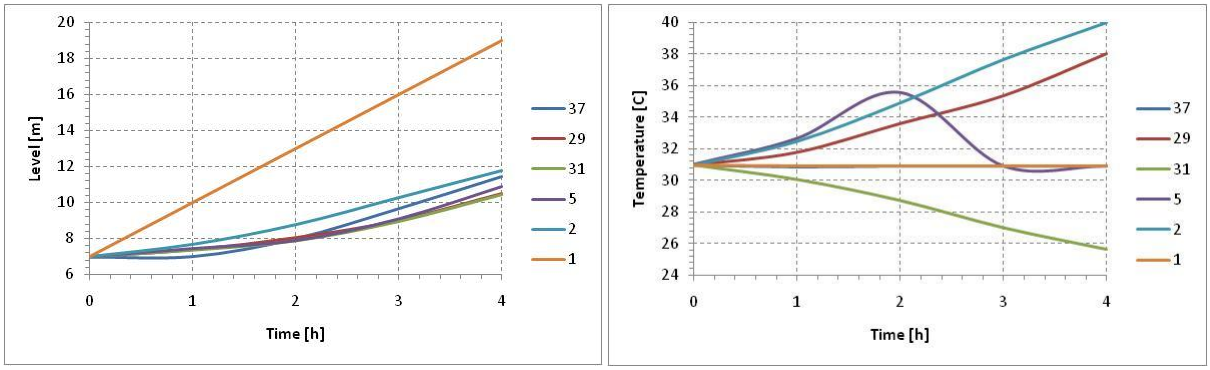


Figure 5: Cluster centers for over-flow (BW=5)

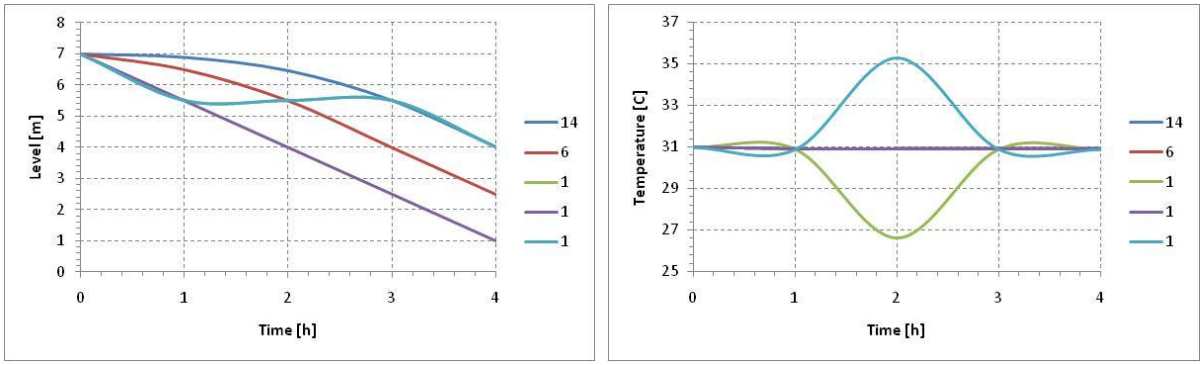


Figure 6: Cluster centers for dry-out (BW=2)

6. CONCLUSIONS

In this paper, we presented a methodology to analyze the set of scenarios generated by the DET methodology. An approach based on the MSM has been presented in order to find the cluster centers. We also coupled the MSM with PCA analysis as a pre-processing tool in order to find the optimal representation of the raw data. We applied this methodology to a set of DETs generated for a simple level controller. The MSM allowed us to identify cluster centers for the both dry-out and overflow Top Events.

7. ACKNOWLEDGMENTS

This work is a product of the project "Risk-Informed Balancing of Safety, Non-Proliferation, and Economics for the Sodium-Cooled Fast Reactor (SFR)" supported by the US Department of Energy under a NERI contract (DE-FG07-07ID14888). The views presented here are those of the authors and do not necessarily represent the views of the US Department of Energy.

References

[1] T. Aldemir, U.S. Nuclear Regulatory Commission, and The Ohio State University, "*NUREG 6901: Current state of reliability modelling methodologies for digital systems and their acceptance criteria for nuclear power plant assessments*", Division of Fuel, Engineering and Radiological Research, Office of Nuclear Regulatory Research, U.S. Nuclear Regulatory Commission, Washington, DC, 2006.

[2] N. Siu, "*Risk assessment for dynamic systems: an overview*", Reliability Engineering and System Safety, vol. 43, no. 1, pp. 43-73, 1994.

[3] G. Cojazzi, "*The Dylam approach for the dynamic reliability analysis of systems*", RESS, no. 52, pp. 279, 1996.

[4] T. Aldemir and M. Belhadj, "*The cell to cell mapping technique and Chapman-Kolmogorov representation of system dynamic*s", Journal of Sound and Vibration, vol. 181, no. 52, pp. 687-707, 1991.

[5] E. Zio D. Mercurio, L. Podofillini and V.N. Dang, "*Identification and classification of dynamic event tree scenarios via possibilistic clustering: Application to a steam generator tube rupture event*", Accident Analysis and Prevention, vol. 41, 2009.

[6] T. Aldemir, "*Utilization of the cell-to-cell mapping technique to construct Markov failure models for process control systems*", in Proc. of Probabilistic Safety Assessment and Management: PSAM1. 1991, pp. 1431-1436, Elsevier, New York.

[7] K. Fukunaga and L. Hostetler, "*The estimation of the gradient of a density function, with applications in pattern recognition*", IEEE Transactions on Information Theory, vol. 21, no. 1, pp. 32-40, 1975.

[8] A. Hakobyan K. Metzroth T. Aldemir R. Denning S. Dunagan B. Rutt, U. Catalyurek and D. Kunsman, "*Distributed dynamic event tree generation for reliability and risk assessment*", in Challenges of Large Applications in Distributed Environments. 2006, pp. 61-70, IEEE.

[9] Y. Cheng, "*Mean shift, mode seeking, and clustering*", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 8, pp. 790-799, 1995.

[10] K. Dubes A. K. Jain and C. Richard, "*Algorithms for clustering data*", Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.