Probabilistic Clustering for Scenario Analysis

D. Mandelli* , K. Metzroth, A. Yilmaz, R. Denning, T. Aldemir
The Ohio State University, Department of Nuclear Engineering, Columbus, OH 43210 U.S.A
** Corresponding author: mandelli.1@osu.edu*

INTRODUCTION

The use of dynamic event trees (DETs) [1] can serve as a powerful tool for the dynamic probabilistic risk assessment (PRA) of nuclear power plants. DETs are similar in structure to their static counterparts [2], except that in DET analysis, time is explicitly modeled and both epistemic and aleatory uncertainties can be accounted for in a phenomenologically consistent manner. The DETs have the capability to more accurately model the complex interactions and events which may occur during a transient.

One of the challenges of dynamic PRA through DETs is the management of the resulting very large data sets. One technique currently being investigated to assist in DET analysis is to group scenarios which have similarities to reduce to number of cases to analyze. An aggregation method which utilizes the Mean-Shift Methodology (MSM) [3, 4] is currently considered for this task. The MSM is a non parametric iterative procedure that shifts each data point to the average of the data points in its neighborhood. The idea behind the MSM is to determine the cluster centers (regions with the highest observation density) and to assign each point to one cluster center only.

What results from the aggregation analysis is a set of representative scenarios with each representative scenario representing a subset of the total scenarios which have similar features. Using this reduced set of representative scenarios, risk results can be constructed without the burden of considering all possible DET cases. The challenge that arises is that using a reduced set of scenarios has the potential to reduce the fidelity of the output risk distribution. In this work, an example case was examined using the ADAPT [5, 6] DET methodology and the risk results were compared between the raw DET data (no scenario aggregation) and the aggregated results using MSM. In addition, a sensitivity analysis was performed on one of the input parameters to the MSM methodology, namely the bandwidth, to determine the sensitivity of the results to this parameter.

DET ANALYSIS

The initiating event investigated was that of a station-blackout (SBO) at a U.S. Pressurized Water Reactor (PWR) and the MELCOR code [7] was linked to the ADAPT tool [6] to determine the evolution for each DET scenario. The simulations using MELCOR as the system code model the transient from the occurrence of the initial condition through the core-melting phase of the accident and up to point of containment failure and release of radionuclides to the environment. For this case, two branching conditions were considered:

1. Creep rupture of major reactor coolant system (RCS) components (hot leg, pressurizer surge line, and steam generator tubes)

2. Failure of the containment vessel

For each of these phenomena, distributions were developed which gave the probability of occurrence as a function of certain system variables. For the first branching condition (creep rupture) a probability distribution was developed [8] on the value of the component lifetime fraction which would lead to failure. For the second branching condition (containment failure), a distribution was developed [8] which gave the probability of containment failure as a function of containment pressure.

Since the ADAPT methodology is discrete in nature, it is necessary to discretize the branching condition distributions. Each branching condition distribution was discretized into seven points, namely, for each of these branching conditions, discrete probability points were selected from the appropriate cumulative distribution functions (CDFs), and the physical values corresponding to these probability values were used as branching criteria. For each branching condition, discrete probability points of 1%, 5%, 25%, 50%, 75%, 95%, and 99% were chosen.

All the 176 scenarios generated in this DET led to containment failure at some point in the scenario evolution. With regards to creep rupture, failure of the surge line dominated this failure mode, with surge line failure occurring also in all scenarios. The DET analysis also showed the potential failure of steam generator tubes before failure of the surge line occurred. However, in this case, steam generator tube rupture was modeled as the failure of a single steam generator tube and this failure did not result in sufficient reactor coolant system depressurization to preclude future failures. As a result steam generator tube rupture had a limited effect on the risk in this scenario as it resulted in little to no containment bypass. Hence, the major contributor to the release of radionuclides to the environment is the over-pressurization and failure of the containment structure (containment failure modes such as basmat melt-through and other energetic containment failure events are not modeled here).

SCENARIO AGGREGATION

The methodology that is presented here is based on the Mean-Shift algorithm which has been described first in [3]. 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

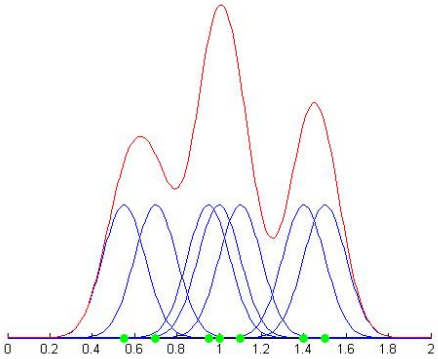

Fig. 1: Density function

cluster center only. By cluster center we mean a region with high observation density (i.e., the modes of the data set).

The main idea is to consider each point $\vec{x}_i (i = 1 \ldots N)$ of the data set as an empirical distribution density function $K(\vec{x}_i)$ distributed in a d-dimensional space (blue line in Fig. 1 for the 1-D case) where regions with high data density (i.e., modes) corresponds to local maxima of the global probability density function $f_N(\vec{x})$ [9, 10] defined as following (red line in Fig. 1 for the 1-D case):

$$f_N(\vec{x}) = \frac{1}{Nh^d} \sum_{i=1}^{N} w_i K\left(\frac{\vec{x} - \vec{x}_i}{h}\right) \tag{1}$$

where each element $\vec{x}_i \in \mathbb{R}^d$ and h is a scalar parameter called bandwidth which indicates the level of refinement of the cluster analysis. The function $K(\vec{x}) : \mathbb{R}^d \to \mathbb{R}$ is the distribution density associated to each data point which is also called *kernel*. In the general formulation of Eq. (1), the terms w_i represent a series of weights that can vary for each data point. In our case, these weights are represented by the probabilities associated with each scenario.

In order to utilize MSM to perform scenario grouping, a set of process variables must be chosen with which to represent the scenarios. For this work, four variables were chosen to represent the state space of each scenario:

1. Average core coolant temperature (x_1)
2. Primary system pressure (x_2)
3. Containment temperature (x_3)
4. Containment pressure (x_4)

Since this scenario deals with both the core melting phase and the containment failure phase of the accident, it was felt that variables should be chosen which represent these phenomena. The first two variables are surrogates for the progression core melt and the integrity of the reactor coolant system, and the second two variables represent the integrity of the containment.

In view of the fact that the system state space consists of 5 variables (i.e., the four variables listed above plus time t), we represent each scenario s_i as a vector in a n-dimensional space as:

$$s_i = [x_1(0), x_2(0), x_3(0), x_4(0), \ldots, x_1(N), x_2(N), x_3(N), x_4(N)] \tag{2}$$

where $x_i(j)$ represents the values of the variable x_i sample at time j. In our simulations, the set of five variables were sampled every 500 seconds (on mission time of $12 \cdot 10^4 s$). Since h is a parameter which must be defined by the user before performing a MSM analysis, we performed the scenario aggregation for various values of h and examined the impact of the choice of h on the fidelity of the risk results.

OBJECTIVE FUNCTIONS

In cluster analysis, one of the frequently occurring dilemmas is: ''how many clusters?'' In numerical analysis a similar question arises when the size of the mesh grid needs to be decided and, in such a case, the objective function is represented by the ''difference'' between the analytical and the numerical solutions.

In our case the objective function F is represented by evaluating the CDF for the containment failure as function of time. In particular, this is accomplished by evaluating the difference between area underneath the CDF for the full data set ($CDF_{FullData}(t)$) and the CDF for the obtained clusters ($CDF_{Clusters}(t)$):

$$F = \left| \frac{\int_0^{12\cdot 10^4} CDF_{FullData}(t)\,dt - \int_0^{12\cdot 10^4} CDF_{Clusters}(t)\,dt}{\int_0^{12\cdot 10^4} CDF_{FullData}(t)\,dt} \right| \tag{3}$$

For our purposes we decided that F should be below 5%; hence, the chosen value of bandwidth h is such that $F \leqq 5\%$.

RESULTS

We performed the MSM analysis for different values of h and compared the cumulative distribution function for containment failures for each case with the raw data. In this respect, Figure 2 shows the CDF for containment failure as a function of time for the raw data set (no aggregation) and for the four different values of h. The timing of containment failure has a major impact on severe accident consequences.

The results show a convergence of the risk results generated from the aggregated data to the raw DET results as the h decreases (see Fig.3). For $h = 14$, the distribution is not well captured in the period before 60,000s. However, the analysis computed with $h = 13$ and $h = 11$ does a much better job of approximating the distribution resulting from the raw data. Finally, the analysis performed using $h = 9$ results in a distribution of containment failure time which closely approximates the distribution resulting from the raw DET data ($F \leqq 5\%$).

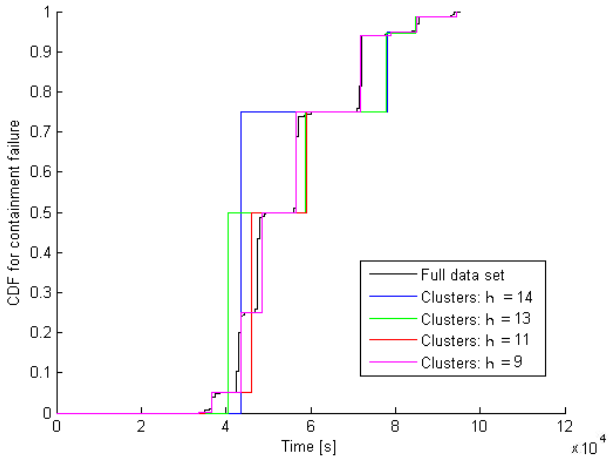

Fig. 2: CDF for containment failure for the whole set of data compared to the ones obtained from the clustering process for different values of bandwidth h

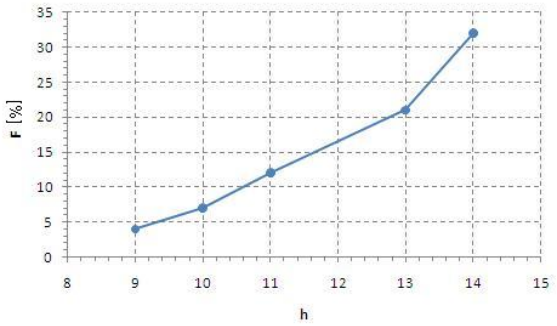

Fig. 3: Objective function F (Eq. 3) as function of the bandwidth h

The aggregation scheme also served to significantly reduce the number of scenarios to be considered. For the case where the $h = 14$, the aggregation scheme produced 4 representative scenarios. Finally, for a $h = 9$, which produced the best characterization containment failure time distribution compared to the raw data, the aggregation scheme produced 15 representative scenarios. As an example, Figure 4 shows a plot of the representative scenarios for the case where $h = 9$. These results imply that the MSM scheme reduced the number of scenarios to consider from 176 down to 15 while still capturing the consequence distribution well.

CONCLUSIONS

This paper presents a study of the Mean Shift Methodology for use in aggregating DET results. In particular, a sensitivity study has been performed using various values of the bandwidth to determine how strongly this parameter impacts the resulting distribution of consequences as compared to the raw, un-aggregated, DET results. The results of this paper show that MSM can significantly reduce the number of scenarios to consider in a DET analysis from 176 to an optimal 15 while still sufficiently capturing the resulting distribution of consequences The massive amounts of data generated in a dynamic PRA represent a major problem for the interpretation of results. An aggregation scheme has been developed here that can reduce to the data to clusters that capture the key features of accident scenarios.

REFERENCES

1. J. DEVOOGHT and C. SMIDTS, ‘‘Probabilistic reactor dynamics. The theory of continuous event trees,’’ *Nuclear Science and Engineering*, **111**, 229--240 (1992).
2. US-NRC, *NUREG 1150 - Severe accident risks: an assessment for five U.S. nuclear power plants*, Division of Systems Research, Office of Nuclear Regulatory Research, U.S. Nuclear Regulatory Commission, Washington, DC (1990).
3. K. FUKUNAGA and L. HOSTETLER, ‘‘The estimation of the gradient of a density function, with applications in pattern recognition,’’ *IEEE Transactions on Information Theory*, **21**, 1, 32--40 (1975).
4. Y. CHENG, ‘‘Mean Shift, Mode Seeking, and Clustering,’’ *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **17**, 8, 790--799 (1995).
5. A. HAKOBYAN, T. ALDEMIR, R. DENNING, S. DUNAGAN, D. KUNSMAN, B. RUTT, and U. CATALYUREK, ‘‘Dynamic generation of accident progression event trees,’’ *Nuclear Engineering and Design*, **238**, 12, 3457 -- 3467 (2008).
6. B. RUTT, U. CATALYUREK, A. HAKOBYAN, K. METZROTH, T. ALDEMIR, R. DENNING, DUNAGAN, and D. KUNSMAN, ‘‘Distributed dynamic event tree generation for reliability and risk assessment,’’ in ‘‘Challenges of Large Applications in Distributed Environments,’’ IEEE (2006), pp. 61--70.
7. R. O. GAUNTT, R. K. COLE, S. A. HODGE, S. B. RODRIGUEZ, R. L. SANDERS, R. C. SMITH, D. S. STUART, R. M. SUMMERS, and M. F. YOUNG, *MELCOR Computer Code Manual, Version 1.8.5, Vol. 2, Rev. 2*, Sandia National Laboratories, NUREG/CR-6119 (1997).
8. A. HAKOBYAN, *Severe Accident Analysis Using Dynamic Accident Progression Event Trees*, Ph.D. thesis, The Ohio State University (2006).
9. T. CACOULLOS, ‘‘Estimation of a multivariate density,’’ *Annals of the Institute of Statistical Mathematics*, **18**, 1, 179--189 (1966).
10. E. PARZEN, ‘‘On the estimation of a probability density function and mode,’’ *The Annals of Mathematical Statistics*, **33**, 3, 1065--1076 (1962).

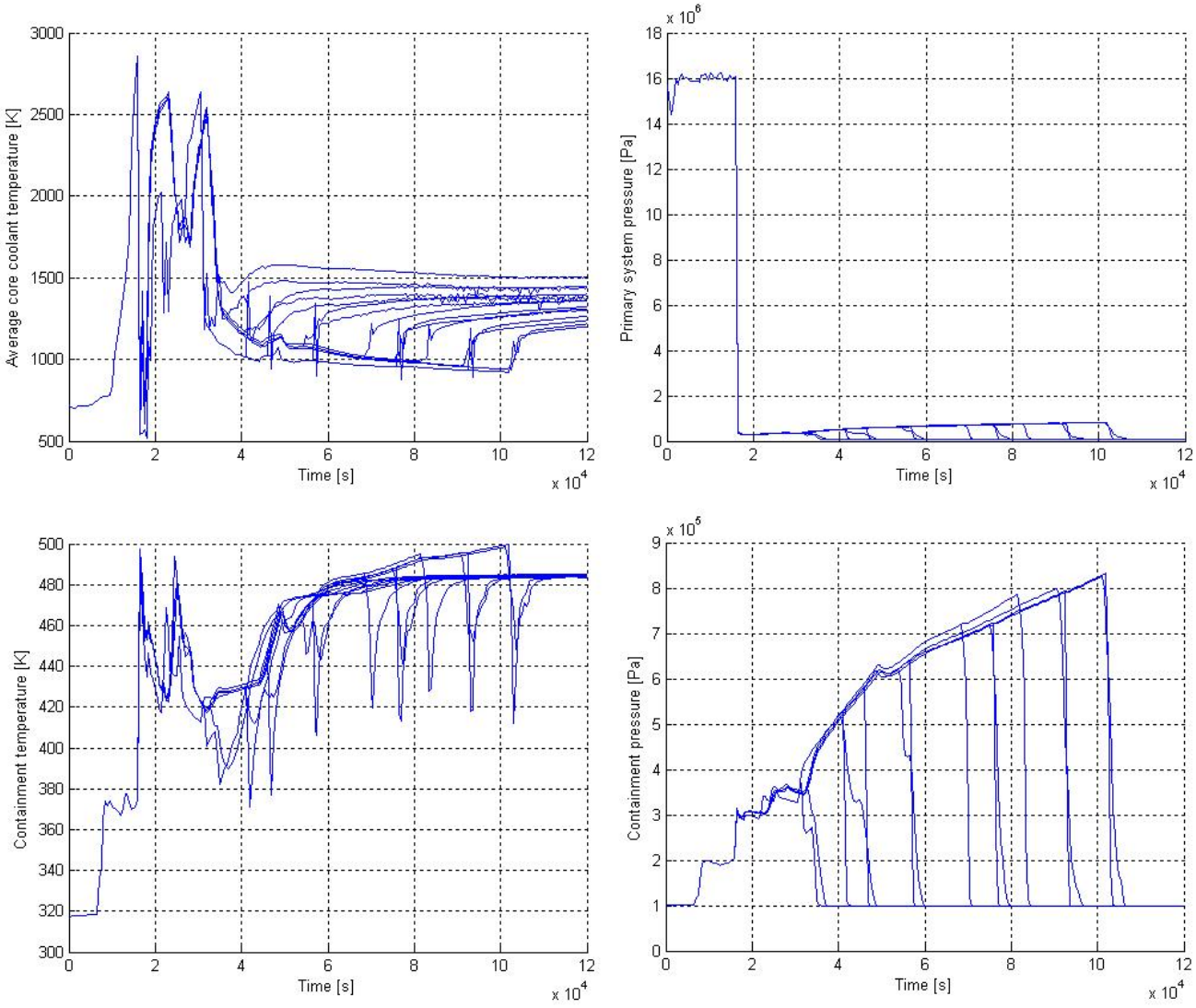

Fig. 4: Cluster centers obtained for $h = 9$