Dynamic PRA: an Overview of New Algorithms to Generate, Analyze and Visualize Data

D. Mandelli, C. Smith, C. Rabiti, A. Alfonsi, R. Youngblood
Idaho National Laboratory, 2525 North Fremont Street, Idaho Falls (*ID*)

V. Pascucci, B. Wang, D. Maljovec
Scientific Computing and Imaging Institute, University of Utah, 72 Central Campus Drive, Salt Lake City (*UT*)

P. T. Bremer
Lawrence Livermore National Laboratory, 7000 East Ave B663-L72, Livermore, (*CA*)

T. Aldemir, A. Yilmaz, D. Zamalieva
The Ohio State University, 281 W. Lane Ave., Columbus (*OH*)

INTRODUCTION

State of the art PRA methods, i.e. Dynamic PRA (DPRA) methodologies [1], largely employ system simulator codes to accurately model system dynamics. Typically, these system simulator codes (e.g., RELAP5 [2]) are coupled with other codes (e.g., ADAPT [3], RAVEN [4] that monitor and control the simulation. The latter codes, in particular, introduce both deterministic (e.g., system control logic, operating procedures) and stochastic (e.g., component failures, variable uncertainties) elements into the simulation. A typical DPRA analysis is performed by:

1. Sampling values of a set of parameters from the uncertainty space of interest
2. Simulating the system behavior for that specific set of parameter values
3. Analyzing the set of simulation runs
4. Visualizing the correlations between parameter values and simulation outcome

Step 1 is typically performed by randomly sampling from a given distribution (i.e., Monte-Carlo) or selecting such parameter values as inputs from the user (i.e., Dynamic Event Tree [3], DET). In Step 2, a simulation run is performed using the values sampled in Step 1). These values typically affect the timing and sequencing of events that occur during the simulation.

The objective of Step 3 is to identify the correlations between timing and sequencing of events with simulation outcomes (such as maximum core temperature). In a classical PRA (event-tree/fault-tree based) environment, such analysis is performed by observing and ranking the minimal cut sets that contribute to a Top Event (e.g., core damage). In a DPRA environment, however, data generated is more heterogeneous since it consists of both:

- Temporal profiles of state variables
- Timing of specific events.

The visual exploration of such data is a new research topic and it is especially relevant when uncertainty quantification is performed on many parameters for complex systems such as nuclear power plants. Such exploration aims to evaluate impact of uncertainties on simulation outcome (e.g., maximum core temperature).

This paper tackles:

- Step 1: How the data is generated
- Step 3: How the data is analyzed
- Step 4: How the data is visualized

and present state-of-the-art algorithms that have been developed in the past few years with the intent of improving the capabilities of DPRA methodologies.

Such algorithms are the result of a series of collaborations between Idaho National Laboratory (under the Risk Informed Safety Margins Characterization project of the Light Water Reactors Sustainability program), University of Utah, the Ohio State University and Lawrence Livermore National Laboratory.

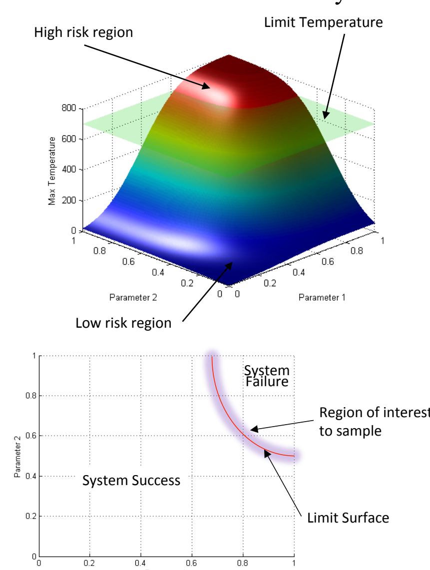

Figure 1. Max core temperature as function of 2 parameters and limit/fail temperature (top) and plot of their intersection: limit surface (bottom)

GENERATE DATA

Nuclear simulations are often computationally expensive, time-consuming, and high-dimensional with respect to the number of input parameters. Thus exploring the space of all possible simulation outcomes is infeasible using finite computing resources. This is a typical context for performing adaptive sampling where a few observations are obtained from the simulation, a surrogate model is built in order predict behavior of the system (e.g., maximum core temperature), and new samples are selected based on the model constructed (see Fig. 1).

The surrogate model is then updated based on the simulation results of the sampled points. In this way, we attempt to gain the most information possible with a small number of carefully selected sampled points, limiting the number of expensive trials needed to understand features of the simulation space. From a safety point of view, we are interested in identifying the limit surface, i.e., the boundaries in the simulation space between system failure and system success. The generic structure of an adaptive sampling algorithm is shown in Fig. 2.

Two classes of algorithms have been evaluated and are being implemented within RAVEN:

- Discrete: model generated predicts simulation outcome in a binary fashion, e.g., system failure or system success
- Continuous: model generated predicts a best estimate of simulation outcome, e.g., maximum temperature reached in the core

In the first class, Support Vector Machines (SVMs) have proven to be flexible to model limit surface of an arbitrary shape [5]. The only limitation is that the surrogate model only predicts the simulation outcome in a binary form (failure or success) and does not give any quantitative information of the variables of interest (e.g., max core temperature). We then investigate algorithms that can generate continuous reduced order models based on Gaussian process models.

We started by evaluating the Kriging method and then developed more advanced algorithms based on topological constructions of the surrogate model (Morse-Smale complexes) [6] as shown in Fig. 3.

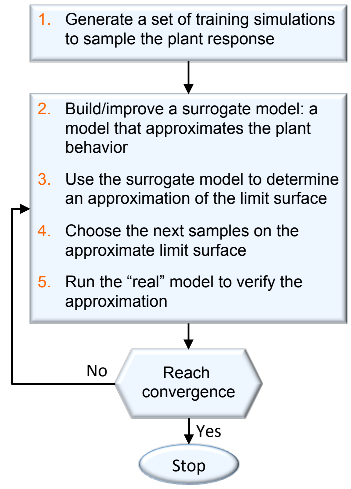

Figure 2. Generic scheme for adaptive sampling algorithms

These algorithms offer better convergence performances, i.e., less samples are need to evaluate limit surfaces. Figure 4 shows an example of limit surface determination for a simplified PWR system during a station blackout (SBO) scenario. Two stochastic variables are considered: initial time after scram (x axis) and duration (y axis) of SBO condition. Note how the uncertainty (green and blue lines) associated to the limit surface (black line) after 10 samples (top of Fig. 4) is very wide while after only 60 samples (bottom of Fig. 4) the limit surface has been completely characterized. Note that

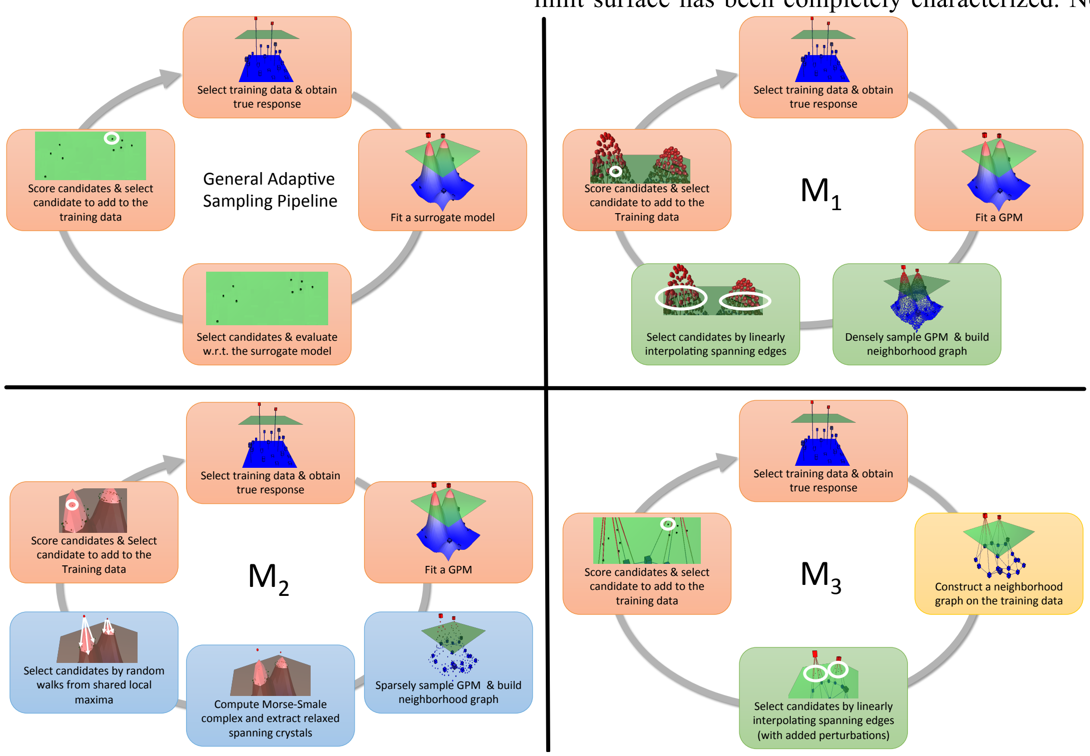

Figure 3. Three topology based methods for adaptive sampling [6]

the limit surface could have been obtained using Monte-Carlo or Latin Hypercube sampling with a much higher number of samples (about 300 samples). Such improvements can be even higher when a large number of stochastic parameters are considered.

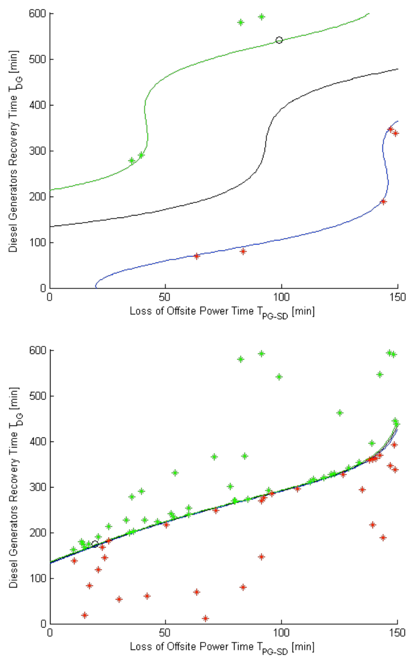

Figure 4. Limit surface obtained for a simplified PWR system for a SBO scenario after 10 (top) and 60 (bottom) samples [5]

While we have primarily testing adaptive sampling schemes to Monte-Carlo analyses, we are also implementing them also to DET analyses.

A slight different approach to select the simulation to run has been shown in [13] and applied to DET analysis. It actually labels scenarios that lead to safe or failure state through a learning process based on Hidden Markov models. The labeling can be applied while the analysis is running and it can be used to select the most significant simulations to runs, and therefore, decrease the execution time of a DET analysis [13].

ANALYZE DATA

The ability to analyze and identify correlations among timing of events through system dynamics/software/human action interactions is essential for nuclear power plant safety analysis and post-processing of the data generated by DPRA methodologies is still a research topic.

A first approach toward discovering these correlations from data generated by DPRA methodologies has been developed using Fuzzy classification. However, clustering algorithms have allowed users to fully analyze these correlations by considering the complete system dynamics and not only the final outcome [7].

Clustering based algorithms can be used to identify groups (i.e., clusters) of scenarios having similar temporal behavior of the state variables. An example [7] is shown in Fig. 5 for a data set generated using ADAPT and RELAP-5 for an aircraft crash scenario. A plot of all 610 scenarios is shown in Fig. 5 (top); clustering algorithm allowed to identify 4 clusters and the “representative scenarios” for each of these 4 clusters are shown in Fig. 5 (middle). At this point, the analysis can be performed by observing the timing of events that lead to the scenarios contained in that cluster (Fig. 5 bottom).

Moreover, clustering algorithms have proven to assist the user, for example, in the identification of those scenarios having similar temporal behavior but characterized by different outcomes only because the maximum simulation time was passed (see Figure 6).

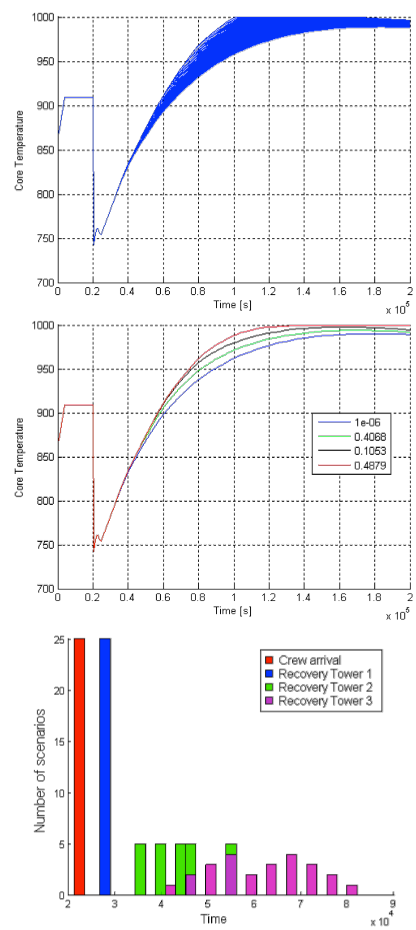

Figure 5. Original data (top), clustered data (middle) and timing of events associated to a cluster (bottom)[7]

In addition, in [7] we showed how clustering algorithms can easily identify outliers scenarios, i.e., scenarios characterized by erroneous/discontinuous temporal behavior for example due to the fact that the validity boundaries of the code were surpassed (see Fig. 7).

In these clustering analyses, only continuous data are used to represent each scenario while discrete data are considered after the clustering process to identify the set of events that caused a similar temporal behavior.

Our recent efforts have been toward the development of methodologies able to analyze scenarios by considering in a coherent fashion both state variables (continuous data) data and timing/sequence of events (discrete data). We accomplished this task by symbolically representing both continuous and discrete datasets [8].

Symbolic representation means that the data are transformed into a series of symbols. Two algorithms are being used:

- A modified version of SAX [9] that discretize state variables symbolically converts (see Fig. 8)
- Time Series Knowledge Representation (TSKR) [10] which symbolically converts discrete types of data (see Fig. 9).

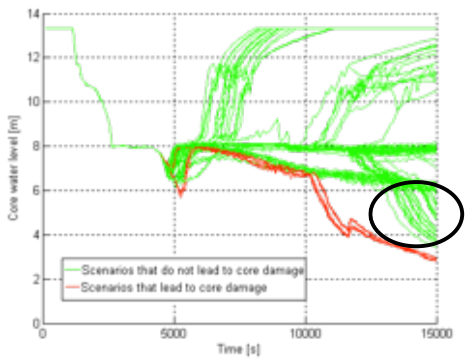

Figure 6. Identification of scenarios that would lead to failure if max simulation time would be extended [7]

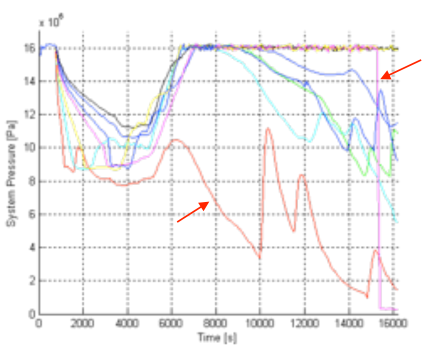

Figure 7 Identification of outliers scenarios generated by erroneopus behavior of the simulation code [7]

These conversions are performed in such way that duration, coincidence and order are preserved. Noteworthy is that high memory requirement reductions were achieved. In addition, we also noticed great computational time reduction when clustering and classification algorithms were applied to the symbolically converted data.

Such reductions (both in term of memory requirements and computational performances) are of relevance for diagnosis and prognosis methodologies when real-time measurements need to be continuously compared with sets of archived data (either generated by simulators or previously monitored and stored).

VISUALIZE DATA

The need for software tools able to both analyze and visualize large amount of data generated by Dynamic PRA methodologies has been emerging only in recent years. In the past 2 years, INL and University of Utah have developed a software tool able to analyze multi-dimensional data: HDViz [11,12].

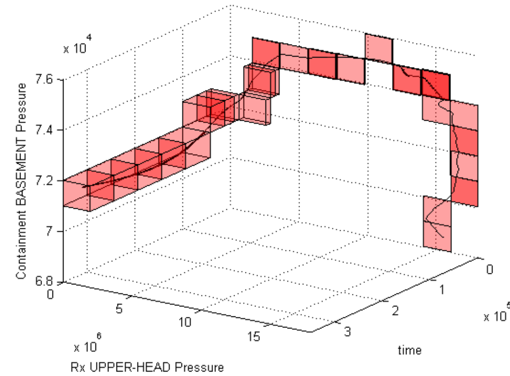

Figure 8. Discretization of a scenario charaterized by two state variables; a specific symbol is associated to each cell [8]

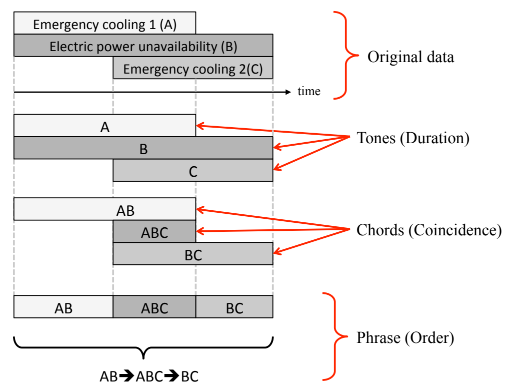

Figure 9. Symbolic conversion of time dependent events [8]

HDViz model the relations between output variables and stochastic/uncertain parameters as high-dimensional functions. In this respect, HDViz segments the domain of these high-dimension functions into regions of uniform gradient flow by decomposing the data based on its approximate Morse-Smale complex (see Fig. 10).

Points (i.e., simulation runs) belonging to a particular segmentation have similar geometric and topological properties, and from these it is possible to create compact statistical summaries of each segmentation.

Such summaries are then presented to the user in an intuitive manner that highlights features of the dataset which are otherwise hidden [11, 12] (see Fig. 11). In addition, the visual interfaces provided by the system are highly interactive and tightly integrated, providing users with the ability to explore various aspects of the datasets for both analysis and visualization purposes.

CONCLUSIONS

This paper has shown several methodologies and algorithms that have been developed among national laboratories and academic research centers. These algorithms are now being evaluated and implemented in projects such as the Risk Informed Safety Margin Characterization (RISMC) and a DPRA code under development at INL: RAVEN.

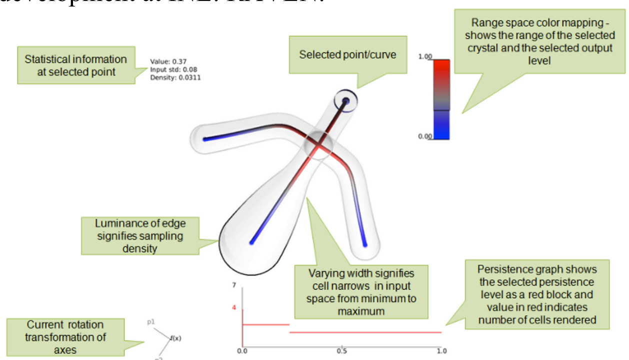

Figure 10. The topological summary visual interface of the simple 2D function [11]

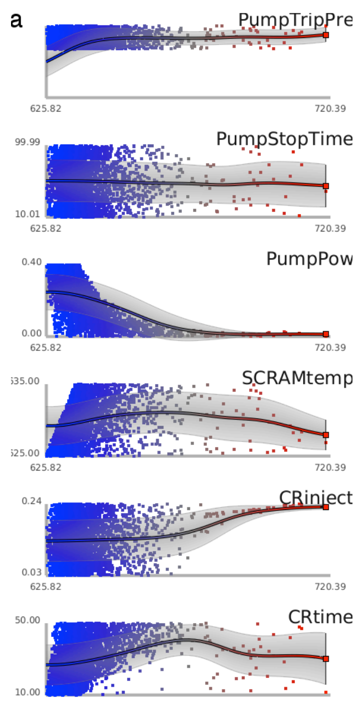

Figure 11. Inverse coordinate plots for a PRA dataset [11]

In this respect, we believe that these algorithms may represent a big step forward toward the utilization of simulation-based methodologies (i.e., DPRA) in order to 1) minimize high computational cost of such analysis (by decreasing the number of scenarios to be generated), and, 2) maximize the amount of information and risk/safety insights that can be explored.

REFERENCES

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

[2] RELAP5-3D Code Development Team, *RELAP5-3D Code Manual*, (2005).

[3] 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**, no. 12, pp. 3457-3467 (2008).

[4] A. Alfonsi, C. Rabiti, D. Mandelli, J. Cogliati, and R. Kinoshita, "Raven as a tool for dynamic probabilistic risk assessment: Software overview," in *Proceeding of M&C2013 International Topical Meeting on Mathematics and Computation,* CD-ROM, American Nuclear Society, LaGrange Park, IL (2013).

[5] D. Mandelli and C. Smith, "Adaptive sampling using support vector machines," in *Proceeding of American Nuclear Society (ANS)*, San Diego (CA), **107**, pp. 736-738 (2012).

[6] Dan Maljovec, Bei Wang, Diego Mandelli, Peer-Timo Bremer and Valerio Pascucci, "Adaptive Sampling Algorithms for Probabilistic Risk Assessment of Nuclear Simulations," draft for *International Topical Meeting on Probabilistic Safety Assessment and Analysis* (PSA) (2013).

[7] D. Mandelli, A. Yilmaz, T. Aldemir, K. Metzroth, and R. Denning, "Scenario clustering and dynamic probabilistic risk assessment," *Reliability Engineering & System Safety*, **115**, pp. 146-160 (2013).

[8] D. Mandelli, C. Smith, T. Aldemir, A. Yilmaz, "Mining nuclear transient data thought symbolic conversion", draft for *International Topical Meeting on Probabilistic Safety Assessment and Analysis* (PSA) (2013).

[9] J. Lin, E. Keogh, S. Lonardi and B. Chiu, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms", in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, San Diego (2003).

[10] F. Morchen, "Time Series Knowledge Mining", *PhD Thesis*, Philipps-University Marburg, Germany (2006).

[11] D. Maljovec, B. Wang, V. Pascucci, P.-T. Bremer, M. Pernice, D. Mandelli, and R. Nourgaliev, "Exploration of high-dimensional scalar function for nuclear reactor safety analysis and visualization," in *Proceeding of M&C2013 International Topical Meeting on Mathematics and Computation*, CD-ROM, American Nuclear Society, LaGrange Park, IL (2013).

[12] D. Maljovec, B. Wang, D. Mandelli, P.-T. Bremer and V. Pascucci, "Analyzing Dynamic Probabilistic Risk Assessment Data through Topology-Based Clustering," draft for *International Topical Meeting on Probabilistic Safety Assessment and Analysis* (PSA) (2013).

[13] D. Zamalieva, A. Yilmaz, T. Aldemir, "Online scenario labeling using a hidden Markov model for assessment of nuclear plant state," *Reliability Engineering and System Safety*, **110**, pp. 1–13 (2013)