Hindawi Publishing Corporation
Science and Technology of Nuclear Installations
Volume 2015, Article ID 278638, 13 pages
http://dx.doi.org/10.1155/2015/278638

Research Article

Scenario Grouping and Classification Methodology for Postprocessing of Data Generated by Integrated Deterministic-Probabilistic Safety Analysis

Sergey Galushin and Pavel Kudinov

KTH, Division of Nuclear Power Safety, AlbaNova University Center, 106 91 Stockholm, Sweden

Correspondence should be addressed to Sergey Galushin; galushin@kth.se

Received 28 January 2015; Revised 27 March 2015; Accepted 31 March 2015

Academic Editor: Francesco Di Maio

Integrated Deterministic-Probabilistic Safety Assessment (IDPSA) combines deterministic model of a nuclear power plant with a method for exploration of the uncertainty space. Huge amount of data is generated in the process of such exploration. It is very difficult to "manually" process and extract from such data information that can be used by a decision maker for risk-informed characterization, understanding, and eventually decision making on improvement of the system safety and performance. Such understanding requires an approach for interpretation, grouping of similar scenario evolutions, and classification of the principal characteristics of the events that contribute to the risk. In this work, we develop an approach for classification and characterization of failure domains. The method is based on scenario grouping, clustering, and application of decision trees for characterization of the influence of timing and order of events. We demonstrate how the proposed approach is used to classify scenarios that are amenable to treatment with Boolean logic in classical Probabilistic Safety Assessment (PSA) from those where timing and order of events determine process evolution and eventually violation of safety criteria. The efficiency of the approach has been verified with application to the SARNET benchmark exercise on the effectiveness of hydrogen management in the containment.

1. Introduction

Development of Deterministic Safety Analysis (DSA) and Probabilistic Safety Analysis (PSA) was crucial step for establishing state-of-the-art in nuclear power safety design and licensing. However, in order to avoid stagnation, it is important to recognize inherent limitations of the classical approaches and new opportunities provided by the overall progress of risk analysis science and computational technologies. For instance, advantage of DSA is that it can model dynamics of the plant systems driven by physical phenomena and their response to failures of the equipment or operator actions. If the "worst" scenarios can be clearly identified, then conservative treatment of uncertainties in DSA can be employed to estimate safety margins. The number of scenarios considered in DSA is usually small with respect to the actual set of possible accident scenarios, thus outcomes of DSA are largely affected by the expert judgment. However, obtaining a priori knowledge about "worst" case scenarios and "conservative" assumptions about uncertain parameters for complex systems is not a trivial task. PSA attempts to cover all possible risk significant scenarios. However, it is not easy to model a priori unknown dependency of the accident scenario outcome on the order and timing of the events (e.g., due to temporary evolution of the system parameters driven by complex physical processes and interactions) using Boolean logic of the classical PSA where the result is unambiguously determined by simple set of events. A robust safety justification must be based on both deterministic and probabilistic considerations to address the effects of the dynamic nature of mutual interactions between (i) stochastic disturbances (e.g., failures of the equipment), (ii) deterministic response of the plant (i.e., transients), (iii) control logic, and (iv) operator actions. Passive safety systems, severe accident, and

containment phenomena are examples of the cases when such dependencies of the accident progression on timing and order of events are especially important. Integrated use of deterministic and probabilistic safety analysis is a means to enable risk-informed decision making based on consistent evaluation of both the uncertainties arising from the stochastic nature of events (aleatory uncertainties) and those arising from lack of knowledge about the processes relevant to the system (epistemic uncertainties) [1].

Integrated Deterministic-Probabilistic Safety Assessment (IDPSA) methodologies aim to achieve completeness and consistency of the analysis through systematic consideration of different sources of uncertainties including physical processes, failures of hardware and software, and human actions. IDPSA tools usually employ (i) system simulation codes and models with explicit consideration of the effect of timing on the interactions between epistemic (modeling) and aleatory (scenario) uncertainties, (ii) a method for exploration of the uncertainty space. A review of the IDPSA methods for nuclear power plant applications can be found in [2].

For decision making, however, it is often insufficient to merely calculate a quantitative measure for the risk and respective uncertainties [3]. Detailed exploration of the uncertainty space usually results in huge amount of the data generated by the deterministic codes [4]. Therefore, one of the main problems for application of IDPSA methods is data post-processing and communication of the analysis results. Extracted information should be suitable for decision making and risk-informed characterization and eventually improvement of safety and performance of the system. Such understanding requires an approach to the interpretation, grouping of similar scenarios, and classification of the principal characteristics of the events that contribute to the risk. Several attempts to solve this problem has been undertaken. Different approaches have been developed to transient identification based on pattern classification by fuzzy C-means clustering [5], identification and classification of dynamic event tree scenarios via possibilistic clustering [6], probabilistic clustering for scenario analysis [7]. These methods use clustering tools and pattern recognition to identify and group similar scenarios that lead to failure.

The goal of this work is to develop methods that will enable understanding of the outcomes of IDPSA analysis while maintaining completeness. In order to achieve that, the methods should reduce the volume of the data generated by IDPSA tools without loss of important for decision making information. The strategy for the reduction of the data volume is based on (i) grouping of different scenarios into different "classes" according to different failure modes; (ii) identification of the scenarios that have "similar" behavior (clustering) within each class. Condensed information should provide useful insights into the complex accident progression and understanding of possible mitigation strategies.

In this work we develop an approach for classification and characterization of failure domains. Failure domain is a domain in the space of uncertain parameters where critical system parameters exceed safety thresholds. The approach is based on scenario grouping and clustering with application of decision trees for characterization of the influence of timing and order of the events. In this approach decision trees are constructed to represent failure domain as a set of leaf nodes and correspondent classification rules that lead to each node. The approach was applied to classification of the simulated transients and failure domain identification and characterization in SARNET benchmark exercise [8].

In this paper we extend our previous work [9] by improving the methods, providing detailed description of the approaches. Specifically, the clustering algorithms and visualization techniques for decision trees have been significantly improved with respect to [9]. In addition, we consider application of developed methods in decision support context.

In Section 2 we provide general description of the approach. In Section 3 we describe a hypothetical accident scenario in a typical French design of Pressurized Water Reactor (PWR). An example of application of the proposed approach to the selected accident scenario is presented in Section 4, followed by the discussion and conclusions.

2. Classification Approach

Methodologies that take into account uncertainty in timing of events can produce potentially unlimited number of transient scenarios for a single initiating event. For decision making, handling of the huge amount of data is a challenge. The development of insights and understanding requires interpretation of the scenario evolutions in order to identify the principal characteristics of the events that contribute to the risk. In order to solve this problem we develop an approach based on clustering and decision trees for explaining the structure of the clustered data (see Figure 1).

The main steps of this approach are briefly explained below. Firstly, the scenario grouping is performed (see Section 2.1). The main idea of this step is to focus the analysis on the sequences intractable in classical PSA. Thus, scenarios where the order and timing of events are not important are grouped first and excluded from further considerations as those directly amenable to PSA analysis. Then we group scenarios where the order of events is important but not their timing. Remaining group of scenarios contains sequences where the outcome depends on the order and timing of the events.

Next, Principal Component Analysis (PCA) [10, 11] is carried out in order identify and quantifying a group of principal components which have the largest influence on the system response (see Section 2.2). Then, based on the PCA results the clustering analysis is performed using Adaptive Mesh Refinement (AMR) method (see Section 2.3.1). In the final step a decision tree is built for each failure mode using clustering results data [12]. Decision tree is used for data representation that explains failure domain-cluster structure (see Section 2.4). The structure is easy to visualize and interpret in the decision-making process. Finally, information of the leaf nodes is used for failure domain probability calculation. Decision tree classification algorithm performs orthogonal partitioning of the search space using data impurity measure as a splitting criterion [10, 13].

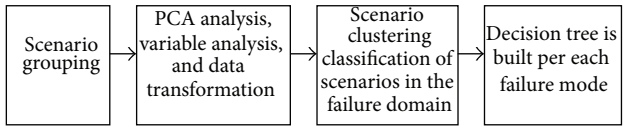


Figure 1: Grouping and classification approach.

2.1. Scenario Grouping. System codes are used in IDPSA in order to evaluate temporal evolution of the accident progression for different time dependent sequences of the events such as activation or failure of safety systems (e.g., reactor protection system and emergency core cooling system). The main purpose of scenario grouping is to identify and separate sequences of events that can be treated in classical PSA, that is, those where order and timing of events have no effect on the outcome (safe or failure end state). The approach is represented in Figure 2.

The numeric algorithm used in scenario grouping is similar to those used in sequence pattern analysis [14]. Each event is represented by a unique number. Thus each simulated transient is represented by a sequence of numbers. Then, for the whole data set, all possible patterns are identified and split into two categories with the same (1) sets of events and (2) order of events. It is important to note that the first category can contain several patterns of the second category (e.g., the set $[2, 3]$ in the first category will represent sequences $(2, 3)$ and $(3, 2)$ in the second category). Then the following steps of the grouping algorithm are performed:

(1) The sets of events that always lead to either failure or safe condition are identified for further treatment in PSA. If the same set of events can lead to both failure and safe states it means that timing and/or order of events can be important. Such sets of events are treated further in Steps (2) and (3).

(2) The sequences of events which always lead to either failure or safe condition are identified. If the same sequence of the events can lead to both failure and safe conditions it is a sign that the influence of timing of the events is important.

(3) The sequences of events where outcome depends on the timing of the events and parameter uncertainty and requires respective dynamic treatment are considered further in the following steps of the analysis, that is, PCA and data transformation, Scenario Clustering, and so forth (see also Figure 1).

2.2. Principal Component Analysis. Principal Component Analysis (PCA) is a technique for revealing the relationships between variables in a data set by identifying and quantifying a group of principal components. These principal components are composed of transformations of specific combinations of input variables that relate to a given output (or target) variable [11]. Each principal component accounts for a decreasing amount of the variations in the raw data set; that is, the first principal component is responsible for the largest possible variance (accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it has to be orthogonal to (i.e., uncorrelated with) the preceding components.

The main purpose of application of PCA in the classification approach is to transform the data without rescaling into a new orthogonal coordinate system that optimally describes the variance in a single dataset. The data transformation is defined by

$$X^{*T} = X^T W, \tag{1}$$

where X^{*T} and X^T are the new and old vectors of observations and W is the matrix of principal component coefficients (eigenvectors of the covariance matrix XX^T) [11].

2.3. Scenario Clustering. The purpose of clustering analysis is to assign members to each group such that members of a group are more similar (according to specific criteria) to each other than to those in other groups (clusters). Clustering analysis is the task of grouping a set of objects in a way that objects within one group (or cluster) are more similar than those in the other groups. It can be achieved by various algorithms that can differ significantly in their notion of what constitutes a cluster and how to efficiently find them. There are several clustering algorithms that methodologically can be separated into connectivity models (hierarchical clustering [15]), centroid based clustering (K-means [15]), distribution based clustering, density based clustering [16], artificial neural networks [17], fuzzy clustering, clustering methodologies based on evolutionary algorithms (Genetic Algorithms [18]), and grid based clustering methodologies [12]. The methodology presented in this paper is based on grid based clustering algorithms with adaptive mesh refinement [12, 19].

2.3.1. Grid Based Clustering. Grid-based clustering methods partition the space into a finite number of cells that form a grid structure on which all of the operations for clustering are carried out. The main advantage of the approach is its computational efficiency [19–21].

Given a set of n-dimensional data and the input parameter, cell size, the search space is partitioned into nonoverlapping rectangular n-dimensional units (cells) of the size ξ. For the sake of conservatism we do not use density threshold for the unit's selectivity parameter (amount of scenarios contained in the unit). Although it might be used in the future development with adaptation of adaptive mesh refinement (AMR) algorithms under conservatism constraints no failure scenarios can be identified as an outlier [19].

Once grid is defined, the algorithm looks for the clusters of cells that contain failure scenarios of the same failure mode. Two cells can form a cluster if they have a common face. The algorithm presents large amount of scenarios with different failure modes as a finite number of cells grouped into clusters corresponding to the same failure mode.

Mesh Refinement. In the adaptive mesh refinement technique the algorithm starts with initial coarse grid. Then, the algorithm identifies the regions with transition between "safe"

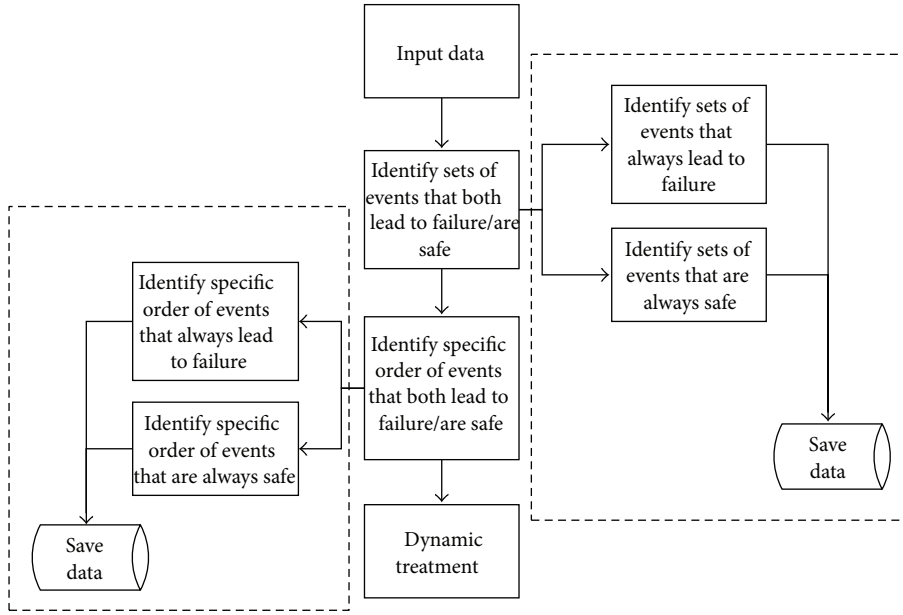


Figure 2: Scenario grouping algorithm.

and "failure" and introduces higher resolution subgrids only in those regions. Finer subgrids are added recursively until either a given maximum level of refinement is reached or the local resolution criterion for the boundary between "safe" and "failure" regions is achieved. Thus in an adaptive mesh refinement computation grid spacing is fixed for the base grid only and is determined locally for the subgrids according to the requirements of the problem.

2.4. Application of Decision Trees. A grid based clustering algorithm performs orthogonal partitioning of the uncertainty space, similar to the partitioning of learning data set in the decision tree. Therefore, complexity of the decision trees can significantly reduce when using clustering results data rather than row scenario data.

A decision tree is a classification and data-mining tool for extraction of useful information contained in large data sets. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute in the given example. This process is then repeated recursively for the subtree rooted at the new nodes until no further branching in the tree can be made or some stopping preset conditions are met [10, 13]. A flow-chart-like structure is generated in which internal nodes represent test on an attribute, each branch represents outcome of test and each leaf node represents class label (decision taken after computing all attributes). Decision trees can be used as a powerful visual and analytical decision support tool; especially in case of multidimensional data, visualization of results in the original space is nontrivial. Decision tree can be constructed using different data impurity measures (e.g., Gini impurity measure and information gain measure) to select the best split among the candidate attributes at each step while growing the tree [13]. Decision trees also can be used as a predictive model which maps observations about an item to conclusions about the item's target value.

2.4.1. Classification and Regression Decision Trees. Most algorithms that have been developed for learning decision trees are variations on a core algorithm that employs a top-down, greedy search through the space of possible decision trees [10, 22]. The best split is identified by a splitting criterion that uses different data impurity measures (e.g., Gini impurity and information gain measure). In this work we use Classification and Regression Tree (CART) with Gini criterion. CART is a nonparametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively [23].

The Gini impurity index (commonly used in CART) at node t is defined as

$$\text{Gini}(t) = \sum_{j \neq i} p(j \mid t)\, p(i \mid t), \tag{2}$$

where i and j are the categories of the target variable, $p(j,t)$ and $p(i,t)$ are proportion of cases in node t with attributes i and j, respectively. Thus, when the cases in a node are evenly distributed across the target categories, the Gini index takes its maximum value $1 - 1/k$, where k is the number of categories for the target variable. The minimum value is zero and it occurs when all the data at a node belongs to one target category.

The Gini criterion for split at s at a node t is defined as

$$\text{Gini}_{\text{split}}(s,t) = \text{Gini}(t) - p_L \text{Gini}(t_L) - p_R \text{Gini}(t_R), \tag{3}$$

where p_L is the proportion of cases in t sent to the left child node and p_R is the proportion of cases in t sent to the right child node. $s \in S$ refers to a particular generic split among all possible sets of splits S.

The split s is chosen to maximize the value of $\text{Gini}_{\text{split}}(s,t)$. Since $\text{Gini}(t)$ is constant for any split s on node t, it can be alternately said that the split s is to be chosen such that the quantity

$$\text{Gain}\,(s,t) = p_L \text{Gini}\,(t_L) + p_R \text{Gini}\,(t_R) \tag{4}$$

is minimized [23].

2.4.2. Probability Estimation Using Decision Trees. The failure domain is represented by agglomerations (clusters) of nonoverlapping cells (grids) in the uncertainty space. If all points in the uncertainty space are equally probable then the probability of the failure domain is the ration of the volume of the failure domain to the total volume of the uncertainty space.

Decision tree represents the failure domain by final nodes in the tree and respective classification rules that lead to these nodes. The probability of each cell can be obtained as average probability of scenarios contained in correspondent cell:

$$\overline{p}_k = \frac{\sum_{i=1}^{N_{\text{scen}}} p_i}{N_{\text{scen}}} \tag{5}$$

and the probability of a failure mode i is

$$p_i = \sum_{j=1}^{N} \sum_{k=1}^{M_j} \overline{p}_k \xi^n, \tag{6}$$

where n is dimensionality, ξ^n is cell volume, $\overline{p}_k$ is average probability of scenarios contained in cell k, M_j are cells contained in the final failure node (leaf) j, and N is total amount of failure nodes (leafs). Depending on the values $\overline{p}_k$ it is possible to assign weights per each cell when building a tree, so the scenarios (cells) with higher probability are likely to be classified into the same final node.

3. Application

In order to illustrate proposed approach we chose a benchmark exercise developed in the framework of the SARNET [8].

The exercise is based on a hypothetical accident transient in typical French 900 MWe PWR (3 loops, with Passive Autocatalytic Recombiners, PAR).

The transient description is as follows:

(i) Loss of coolant accident (LOCA) with a 3″-break size on cold leg of Reactor Coolant System (RCS) (INI – initiation event).

(ii) The Safety Injection System (SIS) and Containment Heat Removal System (CHRS or spray system) which are not available until the beginning of core dewatering.

(iii) The steam generators which are available but not used by the operators.

(iv) No water injection (SIS) occurring before core dewatering.

(v) The reactor operating at nominal power before the initiating event.

(vi) The calculated core dewatering occurring at 4080 s (1 h 08 mn); the vessel rupture occurring at 14220 s (3 h 57 min) if no action is undertaken.

During the core degradation phase, the following assumptions are used:

(i) A water injection (SIS) means is available (with an "average" flow rate) and can be used by the operators.

(ii) The spray system (CHRS) is available and can be used by the operators.

(iii) Water injection after the beginning of clad oxidation causes an increase of the hydrogen flow rate towards containment.

(iv) Hydrogen combustions (hereafter called IGNI event) can occur if the containment gas mixture is flammable; recombiners, because of their high temperature, can initiate a combustion; such combustions can be total (all the hydrogen in the containment is burnt) or not.

For the full list of assumptions made in the benchmark exercise see [8]. For determining the limit of inflammability for the gas mixture Shapiro diagram is used (see Figure 3).

Table 1 gives the limit for inflammability in terms of molar fractions of H_2 versus H_2O.

Water Injection. If water injection occurs before total core uncovery (5875 s), it is assumed that little hydrogen is produced and the vessel rupture is avoided. The probability of this scenario is 0.5.

The probability that water injection is available between total core uncovery (5875 s) and vessel rupture (14220 s) is 0.5. The probability of water injection initiation timing is uniformly distributed in the time interval between total core uncovery and vessel rupture.

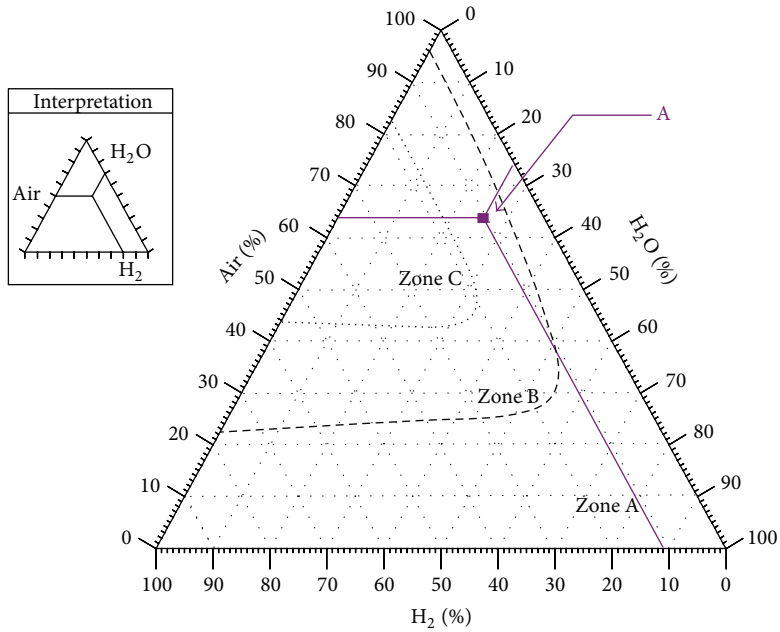
Spray System Activation. The probability that the spray system can be activated after core uncovery (4080 s) and before vessel rupture is equal to 0.5. If the spray system can be activated, the probability of spray system activation is uniformly distributed in the time interval between core uncovery (4080 s) and vessel rupture.

Delay before Combustion. A delay before combustion becomes shorter as H_2 concentration increases. To determine this delay, the following rules are used [8]:

(i) If hydrogen concentration (H_2) < hydrogen inflammability limit (H_{2IF}), no combustion can occur.

(ii) If $H_2 = H_{2IF}$ inflammability limit, the probability of delay before the first (or after previous) combustion is uniformly distributed between 0 and 4 hours.

(iii) If $H_2 \geq$ hydrogen ignition limit (H_{2IG}), the probability of delay before the first (or after previous) combustion is uniformly distributed between 0 and 20 minutes.



Zone A: gas mixture is not flammable; no hydrogen combustion can occur.
Zone B: gas mixture is flammable; hydrogen combustion can occur.
Zone C: it is not considered here.

Figure 3: Shapiro diagram [8].

Table 1: Limit for inflammability.

Molar fraction of H_2O, %	Inflammability limit for H_2 molar fraction, %
0	4
10	4.5
20	5.5
30	6.7
40	8.1
50	10.1

If $H_{2IF} < H_2 < H_{2IG}$, the probability of delay before first (or after previous) combustion is uniformly distributed between 0 and ΔT_{max} (see (7)):

$$\Delta T_{max}(H_2) = \frac{4(H_2 - H_{2IG}) - 0.333(H_2 - H_{2IF})}{H_{2IF} - H_{2IG}}. \quad (7)$$

In this work we consider only containment pressure of $P_{Lim} = 0.3$ MPa threshold as a failure criterion for the sake of simplicity. Using Monte Carlo sampling over 443200 scenarios has been generated for INI (initiating event) + all possible combinations of SIS, CHRS, and IGNI, with different timing of these events.

4. Results

Performing grouping analysis we identified the following possible sequences of the events: [INI SIS]; [INI SIS CHRS]; [INI SIS IGNI]; [INI SIS CHRS IGNI]; [INI CHRS]; [INI CHRS SIS];[INI CHRS IGNI]; [INI CHRS SIS IGNI]; [INI CHRS IGNI SIS]. Classification analysis suggests that sets of events [INI, CHRS], [INI, SIS], [INI, CHRS, SIS], and [INI, SIS, CHRS] do not cause containment over pressurization when they are not followed by hydrogen ignition event (IGNI). Sequences [INI CHRS IGNI] and [INI CHRS IGNI SIS] also do not generate pressure spike big enough to cause containment failure. In the sequences [INI SIS IGNI], [INI SIS CHRS IGNI], and [INI CHRS SIS IGNI] the outcome depends on the timing of ignition (IGNI) and safety systems actuation (see Table 2 for conditional containment failure probabilities for these sequences). In Figure 4 we illustrate an example of application of clustering analysis and decision trees for the sequences that require dynamic treatment.

The advantage of using PCA and coordinate system defined by the principal components of the failure domain is that it significantly reduces the complexity of the decision tree. In case of the transformed coordinate system the decision tree was able to characterize almost 50% of the data set separating the major part of failure scenarios from safe scenarios only in 2 cuts. The results can be transferred back into original coordinate system simply by inverting (1) as follows:

$$X^T = \left(X^{*T} - C\right) W^T, \quad (8)$$

where W is orthogonal matrix ($W^T = W^{-1}$) with principal component coefficients (eigenvectors of the covariance matrix XX^T). In this particular case the values of the W

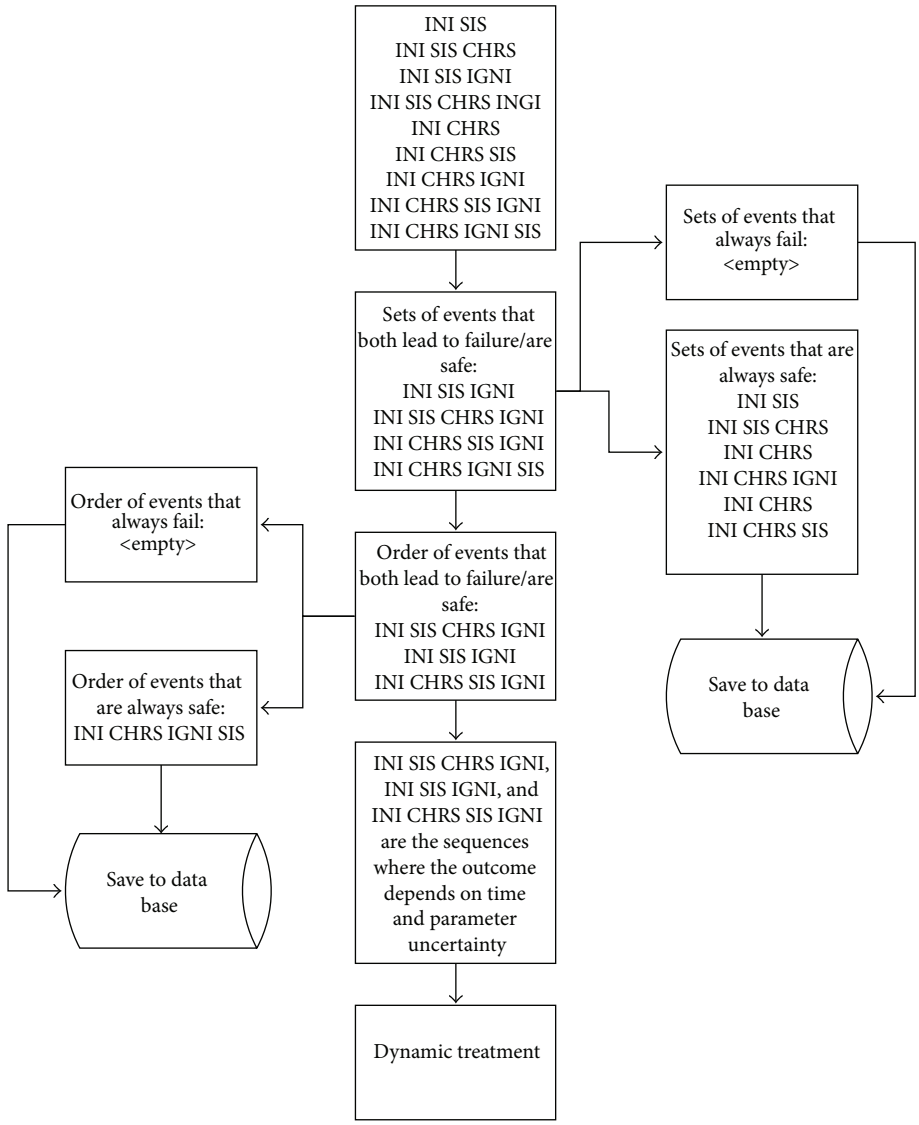


Figure 4: Scenario Grouping.

matrix correspond to ~18.2 degrees rotation counterclockwise, and the variables are defined through the linear combination of variables in original coordinate system:

$$\begin{aligned} \mathrm{SIS}^{*} &= 0.95 * \mathrm{SIS} + 0.31 * \mathrm{IGNI} + 0.31, \\ \mathrm{IGNI}^{*} &= -0.31 * \mathrm{SIS} + 0.95 * \mathrm{IGNI} - 0.007. \end{aligned} \tag{9}$$

The new variables represent linear combinations of all the original parameters involved. The decision tree rules (e.g. $\mathrm{SIS}^{*} > 2955$ sec) in new variables can be also interpreted in the original coordinate system.

Figures 5 and 6 illustrate the results of clustering analysis for the sequence [INI SIS IGNI] with uniform grid. The cells that contain failure scenarios are grouped into cluster representing the failure domain. For each cell in the cluster the algorithm calculates correspondent probability of failure (Figure 7).

Table 2: Containment failure probabilities.

Sequence	Containment failure probability p ($P > P_{\mathrm{Lim}}$)
[INI, SIS, IGNI]	0.51379
[INI, SIS, CHRS, IGNI]	0.07221
[INI, CHRS, SIS, IGNI]	0.00189

Different values of probabilities in the different parts of the failure domain correspond to different H_2 concentrations and respective probability distributions for the time delays of ignition event [8]. For instance, in Figure 8, H_2 concentration is below ignition limit and above inflammability limit; therefore the time delay before the first combustion is uniformly distributed between 0 and $\Delta T_{\max}(H_2)$ (see (7)). In Figure 9, H_2 concentration is above its inflammability and

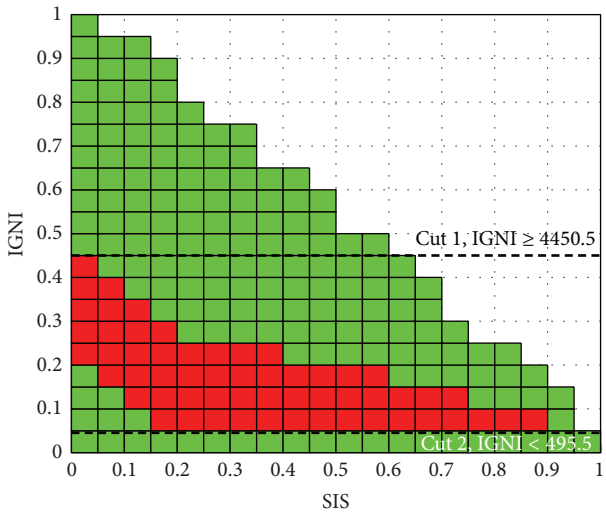


Figure 5: Cluster representation of the failure domain (red) and safety domain (green) for the sequence [INI SIS IGNI], axes scaled between 0 and 1.

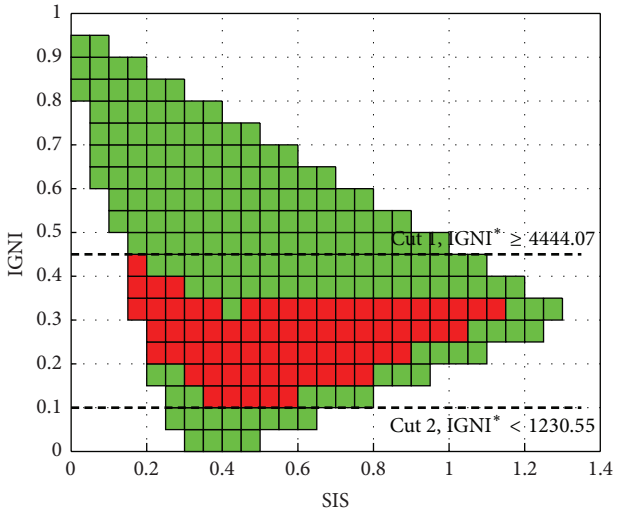


Figure 6: Cluster representation of the failure domain (red) and safety domain (green) for the sequence [INI SIS IGNI] (SIS*, IGNI*: in coordinate system defined by principal components of the dataset), axes scaled between 0 and 1.

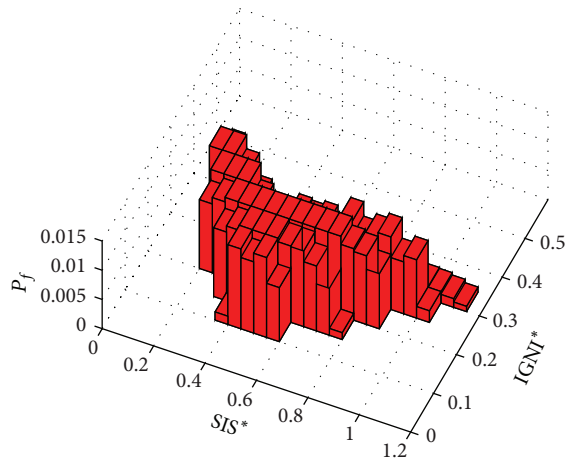


Figure 7: Containment failure probability distribution for sequence [INI SIS IGNI].

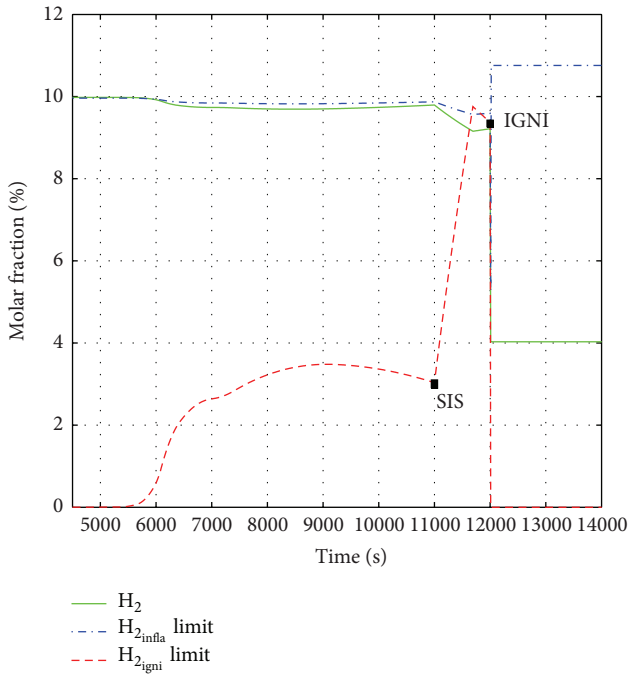


Figure 8: H_2 molar fraction (%) and H_2 inflammability and ignition limits (%).

ignition limits, therefore, according to [8], time delay before combustion is uniformly distributed between 0 and 20 mins.

Failure domain structure can be represented using clustering data and decision tree. To illustrate the approach and to provide a possibility to compare failure domains, presented in Figures 5 and 6, the results are visualized with the decision trees. In this work we use limited amount of uncertain parameters for the sake of visual comparison of the data representation; however, the main advantage of the decision tree approach is the ability to represent complex failure domains with four or more uncertain parameters, when it is difficult to visualize results using other methods. Decision tree complexity depends on the shape of the failure domain and level of details (initial grid and refinement step). However, it is possible to prune decision tree, so the complexity and precision are kept in acceptable levels. Pruning is the process of reducing a tree by turning some branch nodes into leaf nodes and removing the leaf nodes under the original branch [24]. Trees are pruned based on an optimal pruning scheme that first prunes branches giving less improvement in error cost.

After computing an exhaustive tree, the algorithm eliminates nodes that do not contribute to the overall prediction, decided by another essential ingredient, the cost of complexity. This measure is similar to other cost statistics, such as

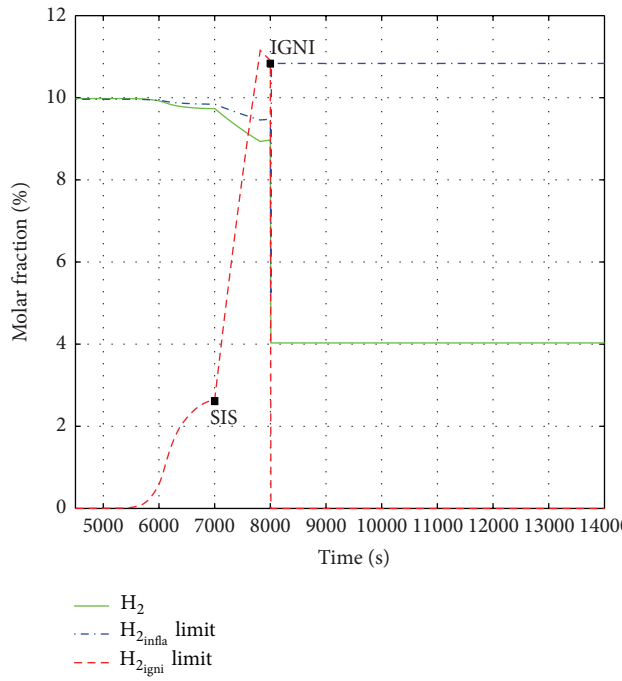


Figure 9: H_2 molar fraction (%) and H_2 inflammability and ignition limits (%).

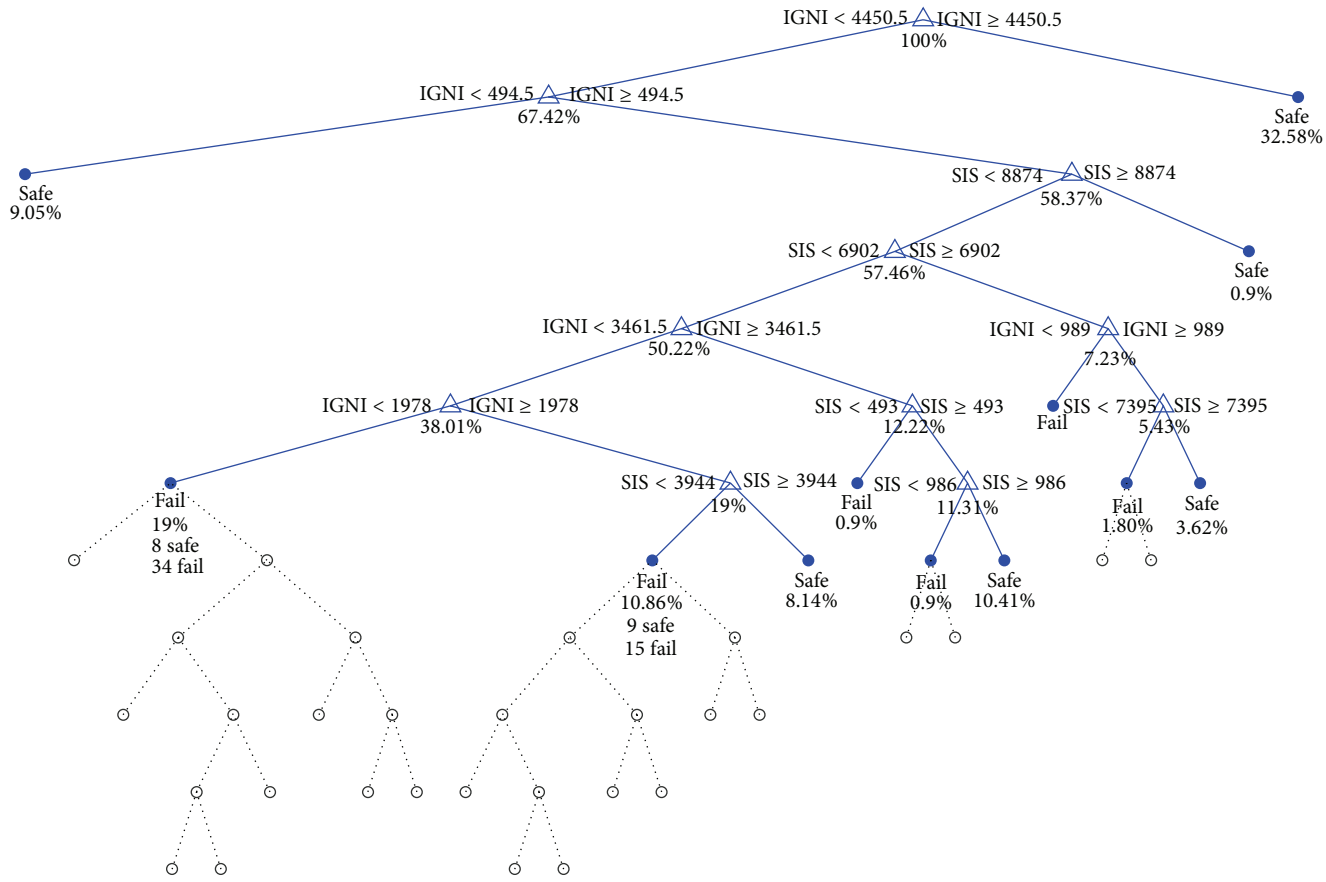


Figure 10: Decision tree fitted into clustering results data for the sequence [INI SIS IGNI] (sec) with pruning.

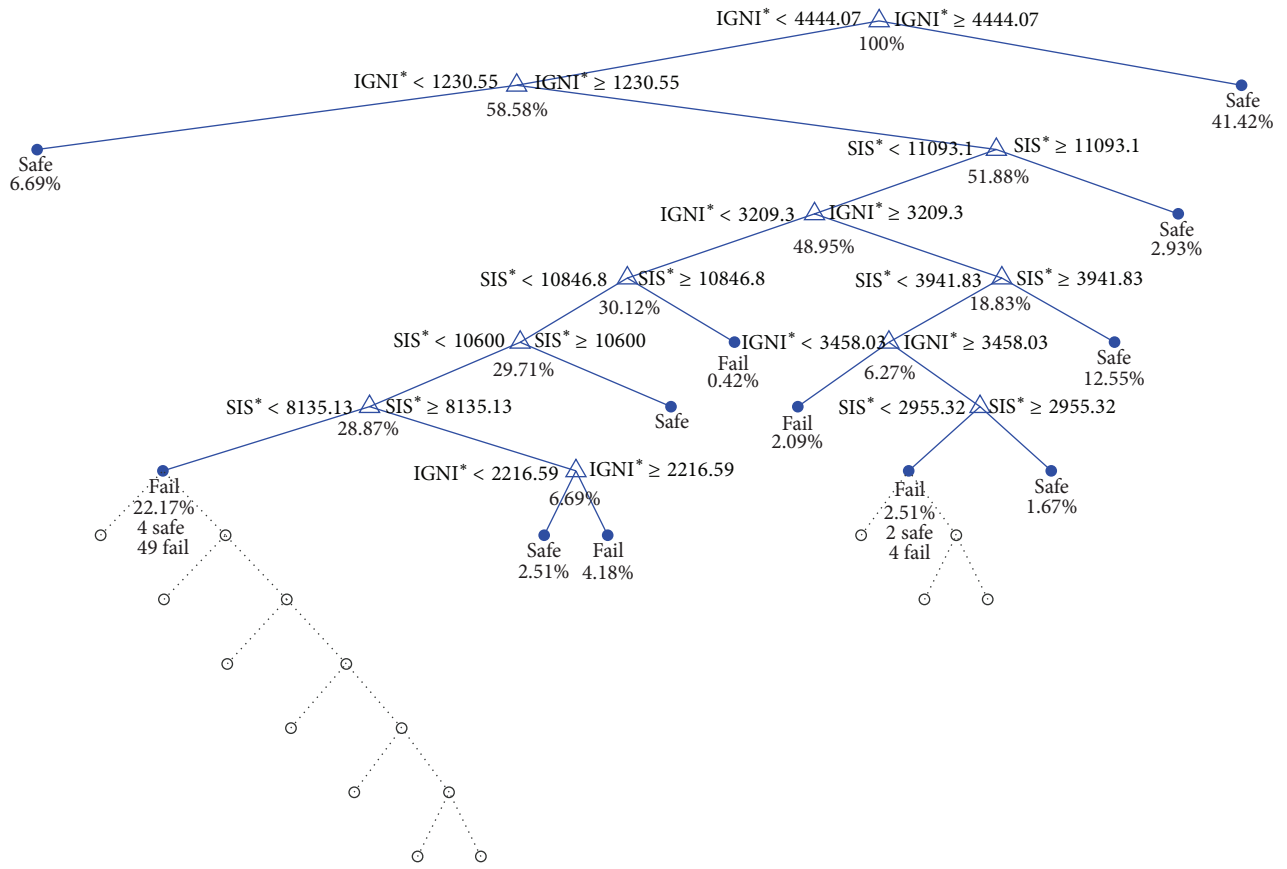


Figure 11: Decision tree fitted into clustering results data for the sequence [INI SIS IGNI] (sec) with pruning (SIS*, IGNI*: in coordinate system defined by principal components of the dataset).

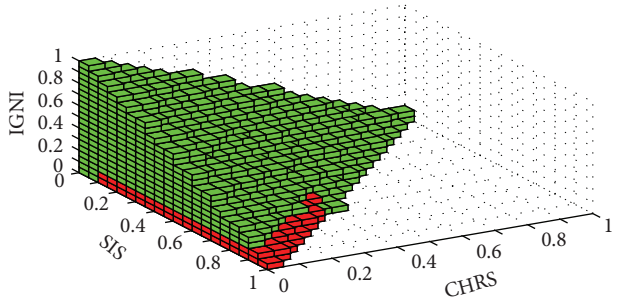


Figure 12: Cluster representation of the failure domain (red) and safety domain (green) for the sequence [INI SIS CHRS IGNI]. Axes scaled between 0 and 1.

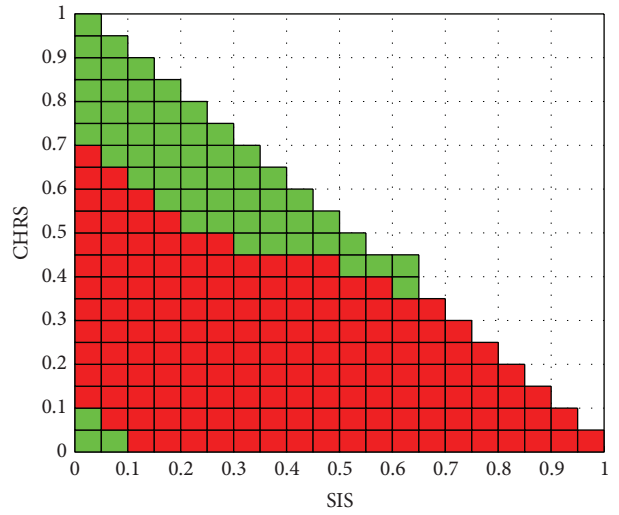


Figure 13: Cluster representation of the failure domain (red) and safety domain (green) for the sequence [INI SIS CHRS IGNI] in terms of controllable events, axes scaled between 0 and 1.

Mallows' C_p [25], which adds a penalty for increasing the number of parameters in a model [24].

Decision tree results for the sequence [INI SIS IGNI] indicate that containment failure is possible if IGNI* event occurs in the time window between 1230.55 and 4444.07 sec (in coordinate system defined by principal components of the dataset). Depending on the timing of the occurrence of the events, H_2 combustion within this time window can challenge containment integrity.

The pruning (cutting) in the decision trees is done at the point where the further refinement will not improve the results and, on the other hand, increase the complexity of the

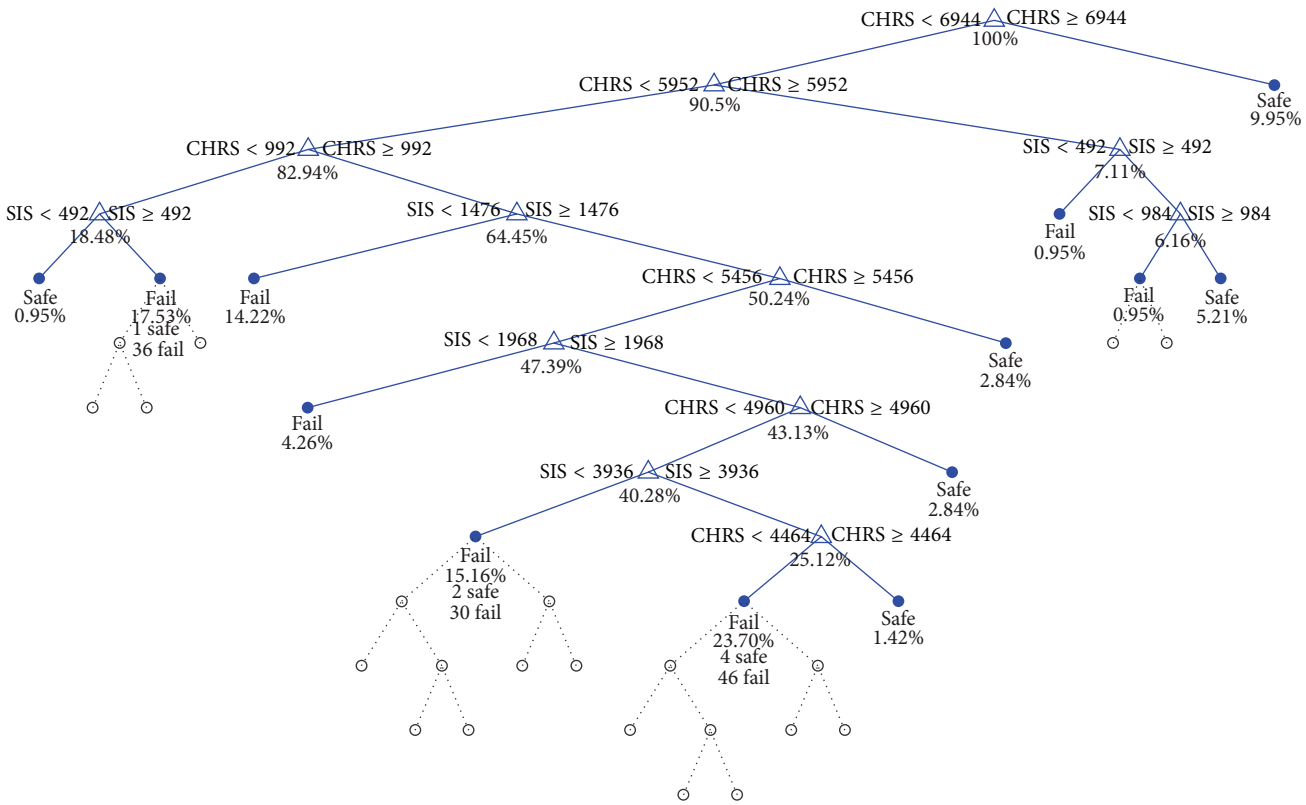


Figure 14: Decision tree fitted into clustering results data for the sequence [INI SIS CHRS IGNI] (sec) for controllable variables.

decision tree. Decision trees (Figures 10 and 11) are built with data set in both original coordinate system and coordinate system defined by its principal components (Figures 5 and 6).

4.1. Decision Support Model. Let us consider as an example of sequence [INI SIS CHRS IGNI]. Figure 12 shows cluster representation of the failure domain in this sequence.

When it comes to decision support, H_2 ignition event (IGNI) in this sequence is entirely stochastic event; that is, the operator has no control over it. On contrary, water injection (SIS) and containment spray (CHRS) systems can be actuated by operator at specified moment of time and, therefore, they are controllable. Decision trees can be used to build decision support model based on the controllable events; that is, decision trees can help us to find an answer to the question "what can be done in case of LOCA initiating event to avoid containment failure?". Figure 13 illustrates failure domain for the sequence [INI SIS CHRS IGNI] in terms of controllable events SIS and CHRS. Based on the clustering results we build a decision tree in variables representing time delays for actuation of the safety systems (SIS and CHRS) and correspondent outcome (Figures 14 and 15). Obtained results indicate that for the sequence [INI SIS CHRS IGNI] containment failure can be avoided in case of early actuation of water injection and containment spray systems (in the range of ~492 seconds) or in case of late activation of containment spray (over ~4000–6944 sec depending on the actuation time of water injection).

5. Discussion

In this work we present an approach for grouping and classification of typical "failure/safe" scenarios identified using IDPSA methods. This approach allows the classification of scenarios that are directly amenable in classical PSA and scenarios where order of events, timing, and parameter uncertainty affect the system evolution and determine violation of safety criteria.

We use grid based clustering with AMR and decision trees for characterization of the failure domain. Clustering analysis is used to represent the failure domain as a finite set of the representative scenarios. Decision trees are used to visualize the structure of the failure domain. Decision trees can be applied to the cases where four or more uncertain parameters are included in the analysis and it is difficult to visualize results in three-dimensional space.

Proposed approach helps to present results of the IDPSA analysis in a transparent and comprehendible form, amenable to consideration in the decision-making process. Useful insights into the complex accident progression logic can be obtained and used for development of understanding and mitigation strategies of the plant accidents including severe accidents. The insights can be employed to reduce unnecessary conservatism and to point out areas with insufficient conservatism in deterministic analysis. Results of the analysis can be also used to facilitate connection between classical PSA and IDPSA analysis.

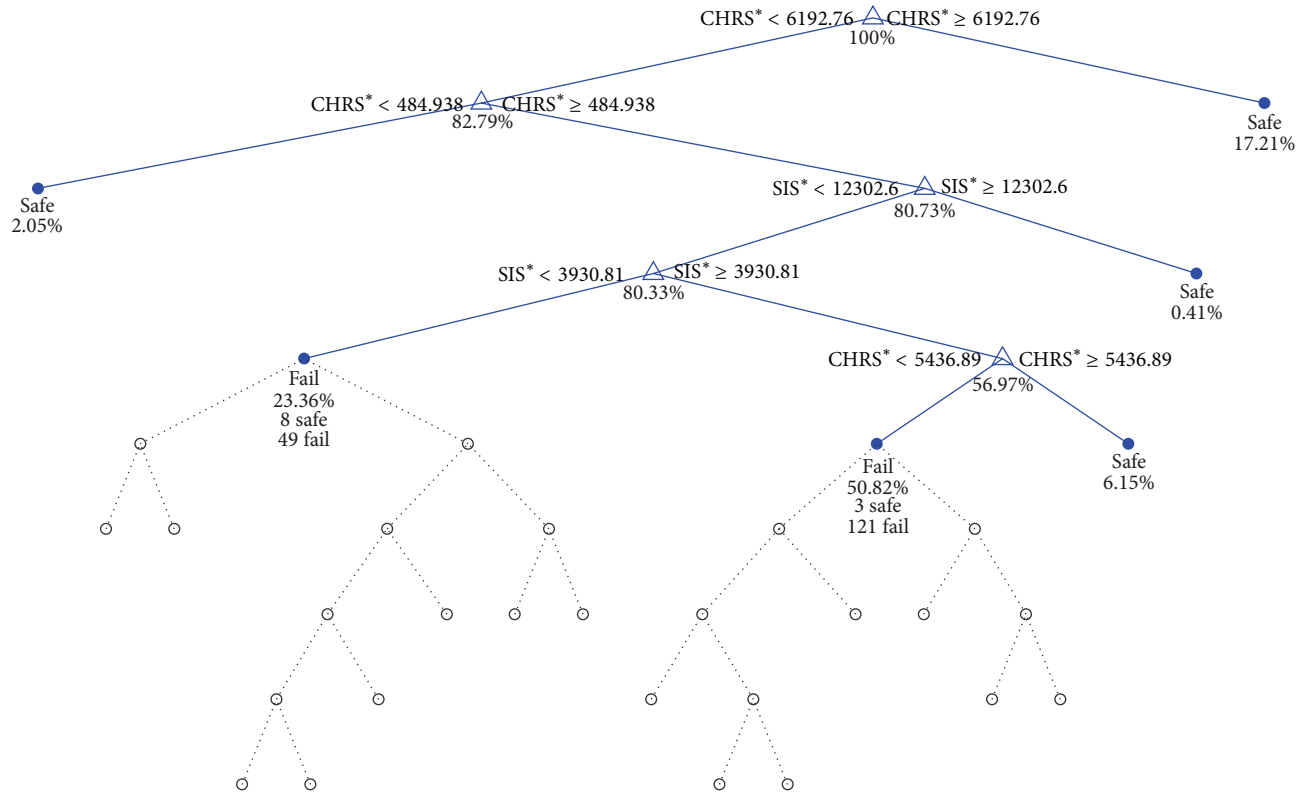


Figure 15: Decision tree fitted into clustering results data for the sequence [INI SIS CHRS IGNI] (sec) for controllable variables (SIS*, CHRS*: in coordinate system defined by principal components of the dataset).

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This study was supported by the Swedish Radiation Safety Authority (SSM). The authors are grateful to Dr. Wiktor Frid (SSM) for very useful discussions.

References

[1] Y. Adolfsson, J.-E. Holmberg, G. Hultqvist, P. Kudinov, and I. Männistö, "Proceedings of the deterministic/probabilistic safety analysis workshop October 2011," Research Report VTT-R-07266-11, VTT, Espoo, Finland, 2011.

[2] T. Aldemir, "A survey of dynamic methodologies for probabilistic safety assessment of nuclear power plants," *Annals of Nuclear Energy*, vol. 52, pp. 113–124, 2013.

[3] S. Hess, "Framework for risk-informed safety margin characterization," EPRI Report 1019206, EPRI, Palo Alto, Calif, USA, 2009.

[4] P. E. Labeau, C. Smidts, and S. Swaminathan, "Dynamic reliability: towards an integrated platform for probabilistic risk assessment," *Reliability Engineering and System Safety*, vol. 68, no. 3, pp. 219–254, 2000.

[5] E. Zio and P. Baraldi, "Identification of nuclear transients via optimized fuzzy clustering," *Annals of Nuclear Energy*, vol. 32, no. 10, pp. 1068–1080, 2005.

[6] D. Mercurio, L. Podofillini, E. Zio, 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, no. 6, pp. 1180–1191, 2009.

[7] D. Mandelli, *Scenario Clustering and Dynamic PRA*, Nuclear Engineering Department, The Ohio State University, 2011.

[8] E. Raimond, "SARNET workpackage 5.3—level 2 PSA Specification of a benchmark exercise relative to hydrogen combustion for application of dynamic reliability methods, IRNS/DSR/SAGR/FT.2005-154," in *Network of Excellence for a Sustainable Integration of European Research on Severe Accident Phenomenology*, IRNS, 2005.

[9] S. Galushin and P. Kudinov, "An approach to grouping and classification of scenarios in integrated deterministic-probabilistic safety analysis," in *Proceedings of the Probabilistic Safety Assessment and Management (PSAM '12)*, Honolulu, HI, USA, June 2014.

[10] S. Tuffery, *Data Mining and Statistics for Decision Making*, Wiley Series in Computational Statistics, John Wiley & Sons, Chichester, UK, 2011.

[11] I. T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, Springer, New York, NY, USA, 2nd edition, 2002.

[12] Ilango and V. Mohan, "A survey of grid based clustering algorithms," *International Journal of Engineering Science and Technology*, vol. 2, no. 8, 2010.

[13] T. M. Mitchell, *Machine Learning*, Mcgraw-Hill Series in Computer Science, McGraw-Hill, New York, NY, USA, 1997.

[14] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data Mining and Knowledge Discovery*, vol. 15, no. 1, pp. 55–86, 2007.

[15] O. Z. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, Springer, New York, NY, USA, 2005.

[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD '96)*, pp. 226–231, AAAI Press, 1996.

[17] L. Fausett, *Fundamentals of Neural Networks Architectures, Algorithms, and Applications*, vol. 16 of *Prentice Hall International Editions*, Prentice Hall, Englewood Cliffs, NJ, USA, 1994.

[18] Y. Vorobyov and T. N. Dinh, "A genetic algorithm-based approach to dynamic PRA simulation," in *Proceedings of the ANS PSA Topical Meeting—Challenges to PSA During the Nuclear Renaissance*, American Nuclear Society, Knoxville, Tenn, USA, 2008.

[19] W.-K. Liao, Y. Liu, and A. Choudhary, "A grid-based clustering algorithm using adaptive mesh refinement," in *Proceedings of the 7th Workshop on Mining Scientific and Engineering Datasets*, Lake Buena Vista, Fla, USA, 2004.

[20] O. Z. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, Springer, New York, NY, USA, 2nd edition, 2005.

[21] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, Calif, USA, 2001.

[22] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers—a survey," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 35, no. 4, pp. 476–487, 2005.

[23] K. P. Soman, S. Diwakar, and V. Ajay, *Data Mining Theory and Practice*, Phi Learning Private Limited, New Delhi, India, 2006.

[24] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, CRC Press, Boca Raton, Fla, USA, 1984.

[25] J. Neter, M. H. Kutner, C. J. Nachtscheim, and W. Wasserman, *Applied Linear Statistical Models*, Irwin, McGraw-Hill, Chicago, Ill, USA, 4th edition, 1996.