Enhanced Power System Security Assessment Through Intelligent Decision Trees

Venkat Krishnan

Abstract Power system security assessment involves ascertaining the post-contingency security status based on the pre-contingency operating conditions. A system operator accomplishes this by the knowledge of critical system attributes which are closely tied to the system security limits. For instance, voltage levels, reactive power reserves, reactive power flows are some of the attributes that drive the voltage stability phenomena, and hence provide easy guidelines for the operators to monitor and maneuver the highly stressed power system to a secure state. With tremendous advancements in computational power and machine learning techniques, there is increased ability to produce security guidelines that are highly accurate and robust under a wide variety of system conditions. Particularly, the decision trees, a data mining tool, has lend itself well in extracting highly useful and succinct knowledge from a very large repository of historical information. The most vital and sensitive part of such a decision tree based security assessment is the stage of training database generation, a computationally intensive process which involves sampling many system operating conditions and performing power system contingency assessment simulations on them. The classification performance of operating guidelines under realistic testing scenarios depend heavily on the quality of the training database used to generate the decision trees. So the primary objective of this chapter is to develop an improvised database generation process that creates a satisfactory training database by sampling the most influential operating conditions from the input operating parameter state space prior to the stage of power system contingency simulation. Embedding such intelligence to the system scenario sampling process enhances the information content in the training database, while minimizing the computing requirements to generate it. This chapter will clearly explain and demonstrate the process of identifying such high information contained sampling space and the advantage of deriving security guidelines from decision trees that exclusively use such an enhanced training database.

Keywords Security assessment · Operating guidelines · Decision trees · Intelligent training set · Monte Carlo simulation · Importance sampling

V. Krishnan (✉)
Department of Electrical and Computer Engineering,
Iowa State University, 1124 Coover Hall, Ames, IA 50011, USA
e-mail: vkrish@iastate.edu

Q. Zhu and A.T. Azar (eds.), *Complex System Modelling and Control Through Intelligent Soft Computations*, Studies in Fuzziness and Soft Computing 319,
DOI 10.1007/978-3-319-12883-2_12

1 Introduction

Traditionally, power system reliability assessments and planning involve deterministic techniques and criteria, which are being used in practical applications even now, such as WECC/NERC disturbance-performance table for transmission planning (WECC 2003; Abed 1999). But the drawback with deterministic criteria is that they do not reflect the stochastic or probabilistic nature of the system in terms of load profiles, component availability, failures etc. (Billinton et al. 1997). Therefore the need to incorporate probabilistic or stochastic techniques to assess power system reliability and obtain suitable indices or guidelines for planning has been recognized by the power system planners and operators; and several such techniques have been developed (Beshir 1999; Chowdhury and Koval 2006; Li and Choudhury 2007; Wan et al. 2000; Xiao and McCalley 2007).

In this regard, Monte Carlo simulation (MCS) methods lend themselves well by simulating the actual analytical process with randomness in system states (Billinton and Li 1994). In this way, several system effects or process including nonelectrical factors such as weather uncertainties can be included in a study based on appropriate parameter's probability distributions. Figure 1 shows an overview of MCS based security assessment methodology, which involves two major tasks: database generation approach and machine learning analysis.

The database generation approach involves the following steps:

- *Random Sampling*: Operating parameters (load, unit availability, circuit outages, etc.) are randomly selected as per a distribution (e.g., uniform, Gaussian, exponential, empirical etc.). This process is generally known as Monte Carlo sampling. Using the generated samples, various base cases are formed.

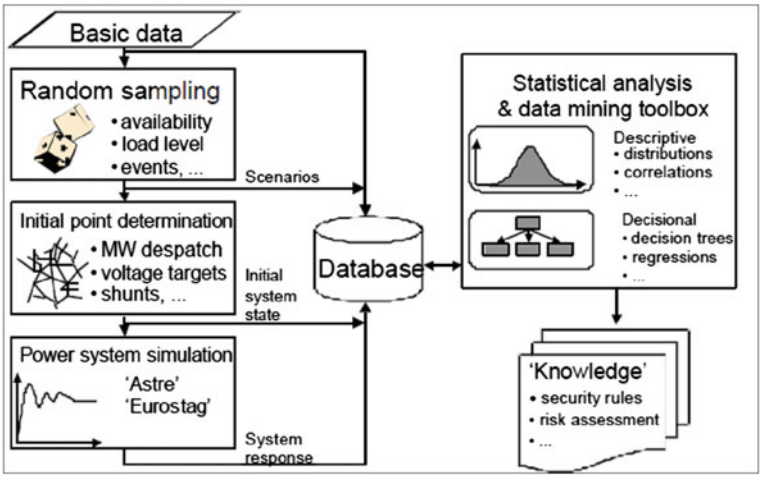

Fig. 1 Probabilistic reliability assessment based on MCS and data mining (Henry et al. 2004b)

- *Optimal power flow*: Initial states for every case is obtained using OPF
- *Contingency assessment*: Using steady-state or time-domain tools contingency events are simulated, and post-contingency performance measures are obtained.

The machine learning methods (Wehenkel 1998; Witten and Frank 2000) are used to extract a high level information, or knowledge from a huge database containing post-contingency responses obtained from the database generation step. These machine learning or data mining techniques are broadly classified as:

- *Unsupervised learning*: Those methods which do not have a class or target attribute. For example, association rule mining can be used to find the correlation between various attributes. Clustering methods such as k-means, EM etc. are generally used to discover classes.
- *Supervised learning*: Those methods that have a class or target attribute, such as classification, numerical prediction etc., and use the other attributes (other observable variables) to classify or predict class values of scenarios. For example, naïve bayes, decision trees, instance based learning, neural network, support vector machine, regression etc.

Among these, decision tree based inductive learning method serves as an attractive option for preventive-control approach in power system security assessment (Zhou et al. 1994; Wehenkel 1997; Zhou and McCalley 1999; Niimura et al. 2004; Wehenkel et al. 2006). It identifies key pre-contingency attributes that influence the post-contingency stability phenomena and provides the corresponding acceptable scenario thresholds. Based on it, security rule or guidelines are developed, which can be deductively applied to ascertain any new pre-contingency scenario's post-contingency performance. Information required for building decision tree are:

- A training set, containing several pre-contingency attributes with known class values
- The classification variable (i.e., class attribute with class values such as "secure" or "insecure"), which could be based on post-contingency performance indices
- An optimal branching rule, i.e., a rule to find critical attribute
- A stopping rule, such as maximum tree length or minimum instances

The aim of inducing a decision tree is to obtain a model that classifies new instances well and produces simple to interpret rules. Ideally we would like to get the best model that has no diversity (impurity), i.e., all instances within every branch of the tree belong to the same class. But due to many other uncertainties or interactions that have not been accounted for in the model, there would be some impurity (i.e., non-homogeneous branch) at most of the levels. So the goal is to select attributes at every level of branching such that impurity is reduced. There are many measures of impurity, which are generally used as optimal branching criteria to select the best attribute for splitting. Some of those are entropy, information gain, Gini index, gain ratio etc.

Classification accuracy and error rates are used as the performance measures of a decision tree. There are two kinds of errors: *false alarms*—acceptable cases classified as unacceptable; and *risks*—unacceptable cases as acceptable. Errors can be calculated by testing the obtained decision model on the training set, which is usually an over-estimate. There are training set sampling methods such as holdout procedures, cross-validation, bootstrap etc. (Witten and Frank 2000) to make the error estimation unbiased. It is even better if the testing is performed using an independent test dataset. There are numerous references that explain the process of building a decision tree from a database with algorithms such as ID3, J48 etc. CART, Answer Tree, Orange, WEKA etc. are some software available for building decision trees.

Many utilities have taken and are continuing to take a serious interest in implementing learning algorithm such as decision tree in their decision making environment. French transmission operator RTE has been using decision tree based security assessment methods to define operational security rules, especially regarding voltage collapse prevention (Lebrevelec et al. 1998, 1999; Schlumberger et al. 1999, 2002; Pierre et al. 1999; Martigne et al. 2001; Paul and Bell 2004; Henry et al. 1999, 2004a, 2006; Cholley et al. 1998). They provide operators a better knowledge of the distance from instability for a post-contingency scenario in terms of pre-contingency conditions, and thus save a great amount of money by preserving the reliability while enabling more informed operational control closer to the stability limits. So the central topic of this chapter will be: *what is the significant component of this decision tree induction process, and how to improve it for the betterment of the planning solutions that are needed under realistic operating conditions?*

The remaining parts of this chapter are organized as follows. Section 2 provides the background of this work in terms of motivation behind this research, related past work, and the objective of this work. Section 3 describes the concept of "information content" in the context of this work. Section 4 presents the technical approach of the proposed high information contained training database generation. Section 5 demonstrates the application in deriving operational rules for voltage stability problem in Brittany region of RTE's system, and presents results and discussions. Section 6 presents conclusions and future research directions.

2 Motivation, Related Work, and Objective

The most vital and sensitive part of MCS based reliability studies is the stage of database generation. The confidence we will have in the results generally reflects the confidence we have in the set of system states generated. The generated database does influence the classification performance of the decision tree against realistic scenarios, selection of critical attributes and their threshold values, and size of the operating rules.

Generally a uniform or random sampling of system states is carried out by varying parameters such as load level, unit availability, exchanges at the boarders, component availability etc. according to their independent probability distributions obtained from projected historical data (Henry et al. 1999, 2004b; Paul and Bell 2004; Lebrevelec et al. 1999; Senroy et al. 2006). Then, various scenarios are simulated for a pre-specified set of contingencies. This stage is generally very tedious and time consuming, as there could be a tremendously large number of combinations of variables [about 5,000–15,000 samples for a statistically valid study (Henry et al. 2004b)]. Therefore, the challenge of producing high information content training database at low computational cost needs to be addressed (Cutsem et al. 1993; Jacquemart et al. 1996; Wehenkel 1997; Dy-Liacco 1997).

In the open literature, there are re sampling methods to retain only the most important instances from an already generated training database (Jiantao et al. 2003; Foody 1999) for classification purposes. But such methods involve huge computational cost in first generating a training database, then identifying the most influential instances, and if need be, generate more of such instances. Genc et al. (2010) proposed an iterative method to a priori identify the most influential region in the operating parameter state space, and then enrich the training database with more instances from the identified high information content region for enhancing classification performance. In this case, the method proposed to identify the high information content region involves heavy computational cost when the dimension of the operating parameter space increases, even beyond 10 parameters.

This chapter proposes to develop an efficient sampling method to generate influential operating conditions that captures high information content for better classification and also reduces computing requirements. In short, the objective is to maximize information content in the training database, while minimizing computing requirements to generate it. This efficient sampling is constructed using the Monte Carlo Variance Reduction (MCVR) techniques. Among the mostly used MCVR methods, control variate and antithetic variate take advantage of the correlation between certain random variables to obtain variance reduction in statistical estimation studies. Stratification method and importance sampling method re-orient the way the random numbers are generated, i.e., alters the sampling distribution (Ripley 1987; Thisted 1988). The proposed efficient sampling method is constructed using the importance sampling method for its ability to bias the Monte Carlo sampling towards the influential region identified a priori; and generate samples within the influential region preserving the original relative likelihood of the operating conditions.

In order to sample the most influential operating conditions, the influential region must be first traced; which requires that the operating parameter state space be characterized with respect to post-contingency performance. A straight forward way to perform state space characterization is to divide the N-dimensional hypercube, where N is the number of selected operating parameters, into M smaller hypercubes, select the center point of each of the M smaller hypercubes and perform an assessment to identify post-contingency performance (NM contingency simulations). But for large N, there is a curse of dimensionality, resulting in very

large computational cost. This work proposes a computationally efficient method based on Latin Hypercube sampling (LHS) to characterize the operational parameter space.

The next section introduces the concept of "high information content" and the measure that can be used to quantify it.

3 High Information Content Region

The decision tree learning algorithm requires a database that has good representation of all the class values, so that it can effectively classify new instances and not overlook the less representative classes. So, for a two-class problem, a good representation of operating conditions on both sides of the class boundary is required. Also, not every operating condition on both sides of the class boundary contributes equally to the operating rule derivation process. This is further demonstrated using Fig. 2 with the help of its four parts a-d, which explain the importance of sampling the most influential operating conditions for the purpose of rule making. For instance, consider sampling some operating conditions defined in terms of variations in Loads A and B as shown in Fig. 2a. Perform contingency analysis to find the post-contingency voltage stability performance (*yellow dots* have acceptable, and *red dots* have unacceptable performances). A suitable rule can be defined by line R that effectively partitions the operating region with acceptable post contingency performance from unacceptable performance. We refer to this line as the security boundary. Now, if more operating conditions are sampled as shown in Fig. 2b, the samples drawn near to the security boundary influences the rule making process more than the samples away from the boundary. This is evident from the consequent rule change (shifting line R) that is necessary as shown in Fig. 2c. So it is very essential that the database contains operating conditions nearer to the security boundary with finer granularity, since they convey more information on the variability of the performance measure, which thereby enables a clear cut decision making on the acceptability of any operating condition. Furthermore, if the some of the operating conditions with unacceptable performance near the rule line R in Fig. 2c are less likely to occur in reality, then the rule line R may be shifted slightly upwards to exploit more operating conditions for economic reasons, as shown in Fig. 2d. Hence the desired influential operating conditions are obtained by sampling according to the probability distribution of the boundary region, which is the shaded region in Fig. 2d where there is a high uncertainty in the acceptability of any operating condition. This will also ensure a very good representation of both the classes in the database at a reduced computational cost compared to sampling from the entire operational parameter state space probability distribution.

In this work Entropy, the most commonly used information theoretic measure for the information contained in a distribution, is used to quantify information content in a database (Unger et al. 1990). It is a function of class proportions, when

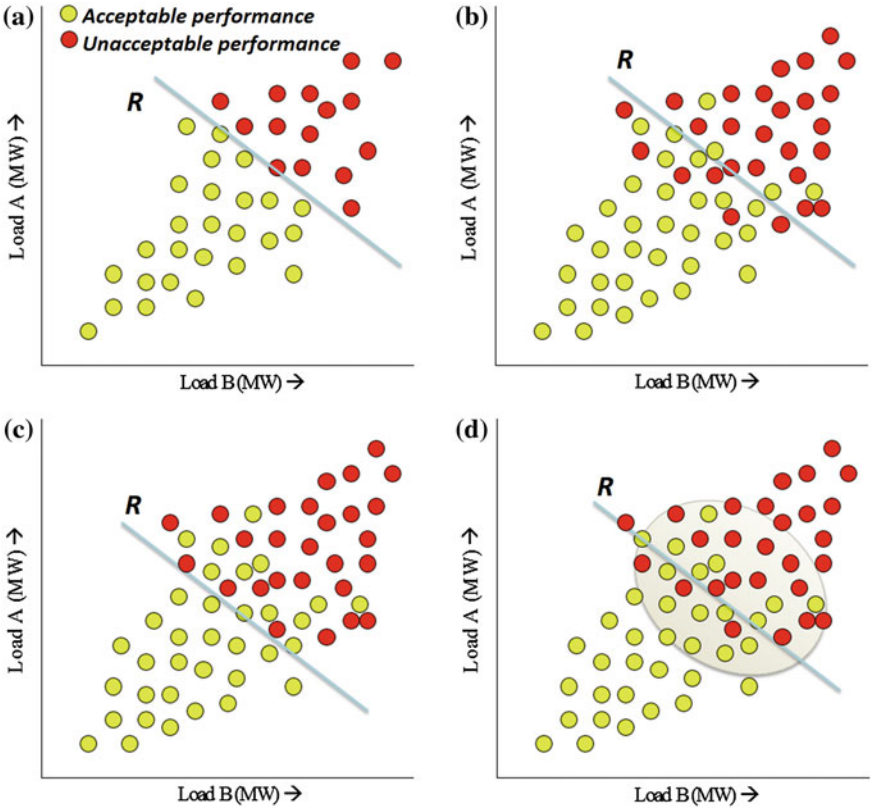

Fig. 2 High information content region

operating conditions are sampled according to their probability distribution. Entropy is given by Eq. (1)

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{1}$$

where, S is training data, c is the number of classes, and pi is the proportion of S classified as class i. Given that the security boundary generally falls in the lower probability region of the operating parameter state space, a database containing samples within the boundary region has the maximum entropy, produced at reduced computational cost. This is the central principle that is used to devise the efficient training database generation approach proposed in this chapter.

The following section will delineate a technical approach that will be used in this chapter to devise the efficient sampling method to generate the high information

contained training database. Later in the numerical results section, the entropy measure introduced in this section will be used to measure the information content in the training database used for producing the decision trees.

4 Technical Approach

The overall flowchart of risk-based planning approach is shown by Fig. 3, along with the proposed efficient sampling approach. The proposed algorithm consists of two stages, where stage I utilizes a form of stratified sampling to approximately identify the boundary region and stage II utilizes importance sampling to bias the sampling towards the boundary region. The database generation is performed for every critical contingency or a group of critical contingencies screened, as depicted by the left-side loop. The right-side loop feeds back information about the region of sampling state space requiring more emphasis in the training database, in order to reduce decision tree misclassifications and improve the accuracy. This chapter primarily focuses on the proposed efficient sampling method.

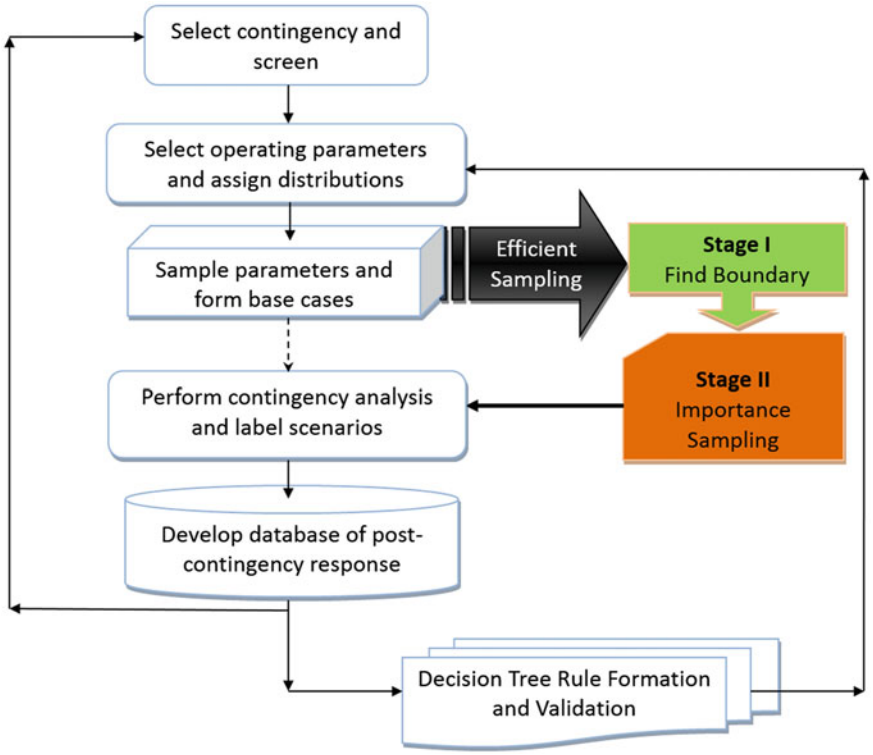

Fig. 3 Proposed approach

4.1 Stage I—Boundary Region Identification

This section develops a LHS method that uses linear sensitivity information to trace the boundary region in a computationally effective manner.

The sampling procedure is computationally very burdensome for a very large dimensional sampling space, especially if the individual load's mutual correlation information is taken into account. So, in order to provide a more reasonable sampling space which would reduce the computation, typically a very strong assumption is made that all loads vary in proportion to the total (also known as homothetic load distribution), so that the load at any bus i maintains a constant percentage of the total load, i.e., $P_{Li} = (P_{Li0}/P_{T0})\,P_T$, where P_{Li0} and P_{T0} are the bus i load and total load in the reference or base case; and P_{Li} and P_T are for any new loading scenario. In the language of voltage stability analysis, these assumptions amount to defining a particular stress direction through the space of possible load increases. Therefore, when a single stress direction is assumed, the uncertainty in load can be simply expressed in terms of the total system load (P_T). So in this case, the sampling is performed only in the univariate space of total system load (P_T) to identify the boundary region.

Generally, this assumption of individual loads having a homothetic distribution along the most probable stress direction is typically done in studies to reduce the computational burden. However, in reality the individual loads may vary along multiple stress directions each having substantial likelihood, and therefore confining to a single stress direction may result in incomplete characterization of the entire load state space. So it is important to consider the multivariate distribution of loads to capture the boundary region effectively. Otherwise, single stress direction assumption will identify only some portion of boundary, and consequently the rules derived from such a database may face challenges when applied to realistic operating conditions. Through the stratified sampling stage (LHS is one kind of stratified sampling), we would want to obtain the boundary region in the multi-dimensional load sampling state space, and then apply the importance sampling to bias the sampling towards this boundary region, which would capture maximum information content including the relative likelihood of sampled operating conditions.

In order to accomplish this, it is necessary to capture inter-load correlations from historical information while sampling from multivariate load distribution to create the training database, where such finer details will have crucial impact in a decision tree's ability to find rules suitable for realistic scenarios. While we can be assured of more information content from this approach, it is likely to increase computing requirements; especially for boundary region identification using stratified sampling. Singh and Mitra (1997) proposed a state space pruning method to identify the important region in a discrete parameter space composed of generation levels and transmission line capacities under a single load level for system adequacy assessment. Yu and Singh (2004) proposed self-organized mapping together with MCS to characterize the transmission line space. Dobson and Lu (2002) proposed a direct and iterative method to find the closest voltage collapse point with reduced computation in the hyperspace defined by loads. But the method's applicability to a

specific distribution of loading conditions in the hyperspace was not shown, and doubts were also cast over its applicability to a large power system with dimension of the hyperspace in 100 s, as will be dealt in the case study of this chapter.

This chapter proposes a sampling space characterization method that uses Latin hypercube sampling (LHS) of homothetic stress directions and linear sensitivities, which promises to reduce the computational requirements. Using this approach the multivariate load state space for a given historical distribution is quickly characterized, under various combinations of Static Var Compensator (SVC) and generator unavailability states. The boundary identification method is described in Sect. 4.1.1, while the stress direction sampling approach (central piece of the proposed state space characterization method) is described in Sect. 4.1.2.

4.1.1 Fast Boundary Region Identification in Multivariate Space

For voltage stability related problems, voltage stability margin (VSM) can be used as the performance measure and hence voltage stability margin sensitivities (Greene et al. 1997; Long and Ajjarapu, 1999; Krishnan et al. 2009) with respect to operational parameters such as individual loads, generator availability, etc. can be used to identify the boundary. VSM is defined as the amount of additional load in a specific pattern of load increase (also termed as stress direction) that would cause voltage instability. It is computed using the continuation power flow (CPF) method. The assumption of a stress direction is important to perform CPF for identifying the voltage collapse point in that direction. Figure 4 depicts existence of several homothetic stress directions for load increase in the two dimensional space defined by loads A and B. The line $Load_A + Load_B = C$ defines various basecases with different inter-node repartitions among loads A and B for the same system load C. These basecases define various homothetic stress directions in the state space, as shown by the various lines from the origin.

The same concept of multiple stress directions is shown in a 3-D load space in the left-hand side of Fig. 5. CPF is performed on these basecases along their

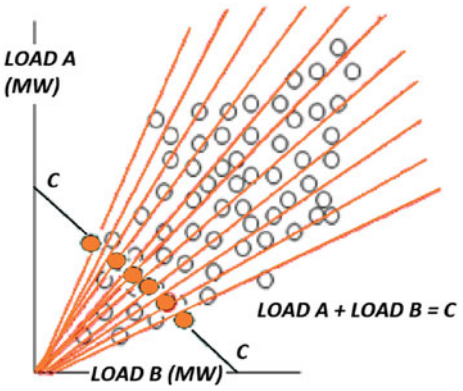

Fig. 4 Multiple homothetic stress directions

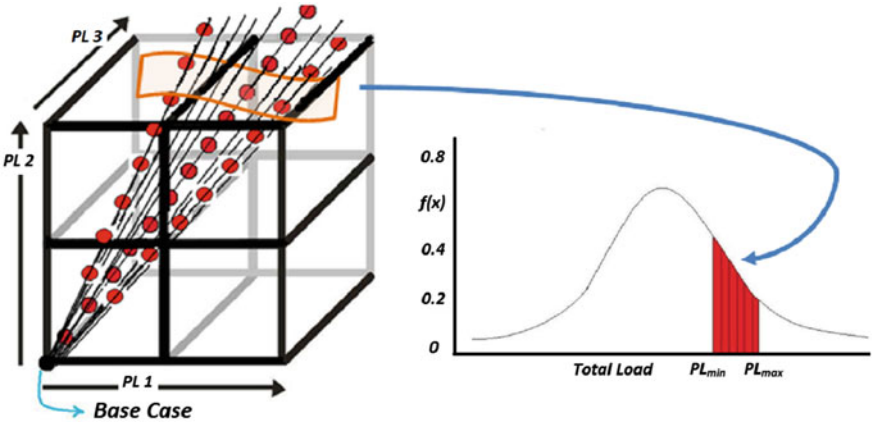

Fig. 5 Multiple homothetic stress directions in 3-D and boundary identification

intrinsic load increase directions, and the maximum loadability along each stress direction is computed. From these the boundary limits, $\{P_{Lmin}, P_{Lmax}\}$, in the total system load space is found, as shown in the right-hand side of Fig. 5. This limit in the hyperspace is subject to variation due to the influence of discrete variables such as SVC and generator unavailability states. The effect of these two variables is estimated using VSM sensitivities with respect to real and reactive power injections along every stress direction, and is given by the Eq. (2). Usage of such linear sensitivities significantly reduces the computational burden in characterizing a multi-dimensional operational parameter state space.

$$\Delta P_L^{svc} = Q_{svc}^{*} \cdot dVSMdQ_{svc} \tag{2}$$

where ΔP_L^{svc} is the change in boundary limit in a particular stress direction due to the influence of SVC unavailability, Q_{svc}^{*} is the amount of unavailable SVC reactive power at the collapse point, and $dVSMdQ_{svc}$ is the linear sensitivity of voltage stability margin with respect to reactive power injection at the SVC node, which is computed as a by-product of CPF study in that particular stress direction.

Finally, the boundary limits in the total system load space is identified, subject to these discrete variable influences. The key in realizing the computational benefit in boundary region identification lies in the manner in which the multiple homothetic stress directions are sampled from the historical data.

4.1.2 Sampling Homothetic Stress Directions Using Latin Hypercube Method

Latin Hypercube Sampling (LHS) is very prevalently used in Monte Carlo based reliability studies in many fields. LHS of multivariate distribution is performed by

dividing every variable forming the multivariate distribution into k equiprobable intervals, and sampling once from each interval of the variable. Then these samples are paired randomly to form k random vectors from the multivariate distribution. Figure 6 depicts the stratified sampling in both forms, traditional and LHS, where the difference is in the pairing process. In the traditional stratified sampling, samples from every interval of variable i is paired with every other samples from all intervals of variable j; whereas in the LHS, one sample from an interval of variable i is paired only once with any one of the sample from an interval of variable j. The pairing in LHS can also be done in such a way as to account for the mutual correlation of the variables by preserving their rank correlation (Wyss and Jorgensen 1998), and hence capturing the inter-dependence structure of the multivariate distribution.

Similarly, LHS of homothetic stress directions is performed from historical data by dividing the load stress factor variables into k equidistant intervals (i.e., equal width; a modification to traditional LHS that partitions into equiprobable intervals), sampling once from each interval of the variable, and pairing them preserving their rank correlation, to form k homothetic stress directions. Figure 7 shows (a) traditional stratified sampling and (b) LHS of homothetic stress directions in 3-dimensional state space. In the case of LHS, for k intervals per dimension, irrespective of state space size the uniform stratification of stress direction is achieved with k samples; compared to the stratified sampling that produces k^{n-1} samples for k intervals per dimension, in a state space of dimension n. The ideal number of k is found in an incremental fashion until there is no further improvement in the boundary limits. Hence computation to find the boundary region can be decreased drastically by using the proposed method based on LHS of stress directions and linear sensitivities.

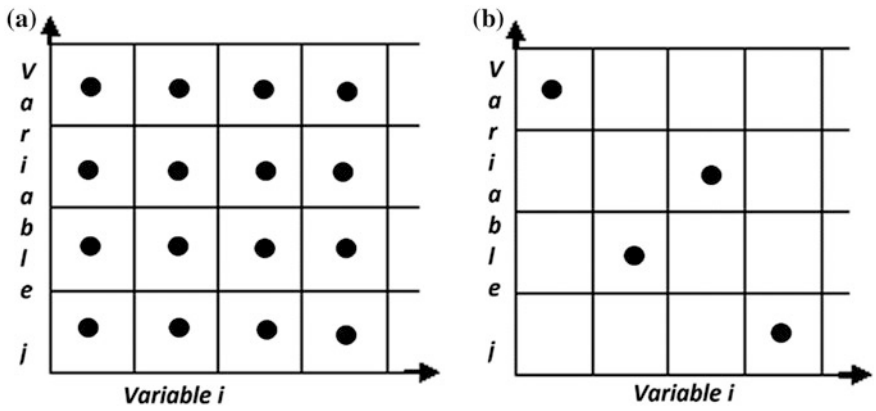

Fig. 6 Stratified sampling—**a** traditional, **b** LHS

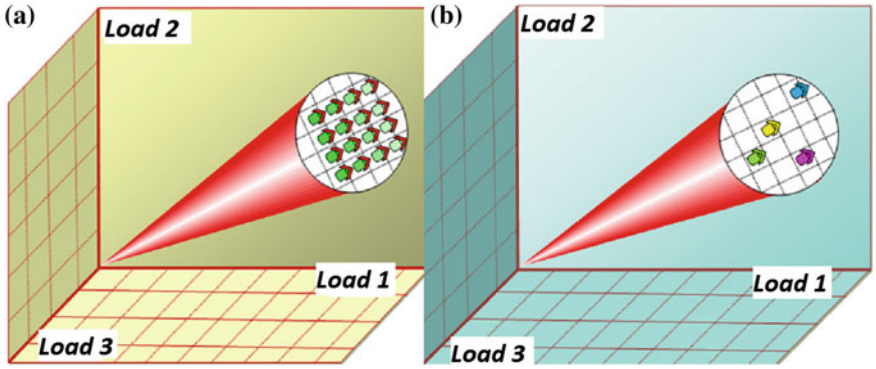

Fig. 7 Sampling homothetic stress directions for boundary identification. **a** Traditional stratified sampling. **b** Latin hypercube sampling

4.2 Stage II—Importance Sampling

Once the boundary region has been identified, the next step is to sample operating conditions from that. This section describes the central concept behind embedding such intelligence in the sampling approach.

The standard Monte Carlo sampling approach draws values for each parameter in proportion to the assigned distribution. Given the previous knowledge of the boundary region from Stage I, biasing the sampling process towards the boundary region can be implemented using the importance sampling method, which helps in maximizing the information content. In this study, the inter-load correlations are captured in the sampling process using copulas (Papaefthymiou and Kurowicka 2009), unlike many studies that approximate the inter-load correlations using multivariate Normal distribution for computational purposes. Copulas are generated based on non-parametric historical load distribution, and it enables sampling realistic scenarios.

4.2.1 Importance Sampling Variance Reduction

In risk-based security planning studies, the quantity of interest is probability of unacceptable performance, i.e., $P(Y \sim \textit{unacceptable events})$ (Billinton and Li 1994).

$$P(Y<t) = \int_{-\infty}^{t} f(y)dy \tag{3}$$

where, $y = t$ denotes the threshold performance such that $y < t$ is unacceptable performance. The indicator function $I(y)$ denoting region of interest $h(y)$ is defined as,

$$h(y) = I(Y < t) = \begin{cases} 1 \text{ if } Y < t \\ 0 \text{ if } Y \geq t \end{cases} \tag{4}$$

and hence,

$$P(Y < t) = \int_{-\infty}^{\infty} h(y) f(y) dy = E(h(Y)) = \sum_{i=1}^{n} h(y_i) \tag{5}$$

The above expectation function gives crude Monte Carlo estimation (Rubinstein 1981), where *yi* are Monte Carlo samples taken from the distribution $f(y)$, the post-contingency performance index probability distribution. This estimation has a variance associated with it, as the quantity $h(y_i)$ varies with y_i. Importance sampling attempts to reduce the variance of the crude Monte Carlo estimator by changing the distribution from which the actual sampling is carried out. Suppose it is possible to find a distribution $g(y)$ such that it is proportional to $h(y)\, f(y)$, then the variance of estimation can be reduced by reformulating the expectation function as,

$$P(Y < t) = \int_{-\infty}^{\infty} h(y) f(y) \frac{g(y)}{g(y)} dy = E\left(\frac{h(Y) f(Y)}{g(Y)}\right) = \sum_{i=1}^{n} \frac{h(y_i) f(y_i)}{g(y_i)} \tag{6}$$

where y_i are Monte Carlo samples drawn from the distribution $g(y)$. This ensures the quantity $\{h(y_i) f(y_i)/g(y_i)\}$ is almost equal for all y_i. In effect, by choosing the sampling distribution $g(y)$ this way, the probability mass is redistributed according to the relative importance of y, measured by the function $|h(y)|\, f(y)$ (Ripley 1987).

4.2.2 Proposed Efficient Sample Generation

The property of importance sampling to bias the sampling using an importance function $g(y)$ towards an area of interest, as discussed above is used to generate influential operating conditions from operational state space, X. The joint probability distribution of the operational parameter space $f(x)$ can be obtained from historical data (Rencher 1995). Once we have *a priori* information about $f(x)$, stage-I operation provides the region in X through which the boundary most likely occurs and therefore identifies approximately the x-space in which we want to bias the sample generation. The region of interest for sampling is defined using the indicator function $h(X)$, where S is the boundary region.

$$h(X) = I(X \in S) = \begin{cases} 1 \text{ if } Y(X) \in S \\ 0 \text{ if } Y(X) \notin S \end{cases} \tag{7}$$

In a univariate case, we can define it as $S = \{x : x_1 \leq x \leq x_2\}$, as shown in Fig. 8. The importance function or the sampling distribution $g(x)$ can be constructed

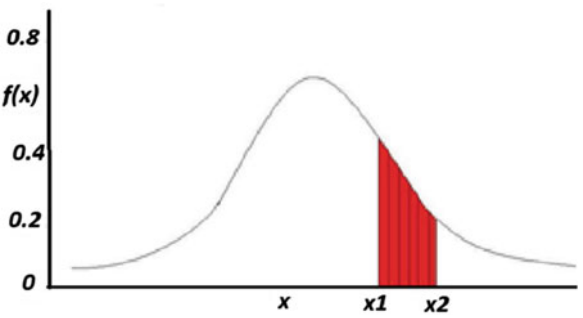

Fig. 8 Boundary region in the univariate operating parameter distribution $f(x)$

proportional to $|h(x)|\ f(x)$, i.e., focusing on the region S of $f(x)$. In general, the importance sampling density can be expressed as,

$$g(x) = p * f_1(x)\ *\ \mathrm{I}(x \in S) + (1-p) * f_2(x)\ *\ \mathrm{I}(x \notin S) \tag{8}$$

where p controls the biasing satisfying the probability condition $p \leq 1$, $f_1(x)$ is the probability density function of the boundary region, and $f_2(x)$ is the probability distribution function of the region outside boundary. We can adopt a composition algorithm to generate samples from the distribution $g(x)$ (Devroye 1986; Gentle 1998). Setting $p = 0.75$, 75 % of the points can be expected from region S, thereby performing an upward scaling of the distribution $f(x)$ towards the boundary region.

In the multivariate case, sampling techniques such as copulas or LHS or sequential conditional marginal sampling (SCMS) (Papaefthymiou and Kurowicka 2009; Hormann et al. 2004) is used to generate correlated multivariate random vectors from non-parametric distributions $f_1(x)$ and $f_2(x)$. The SCMS method is time consuming and requires a lot of memory usage for storing the entire historical data, while LHS and copulas are relatively faster and consume less memory since they work only with non-parametric marginal distributions and correlation data. We use copulas for their simpler and elegant approach in handling any non-parametric marginal distributions

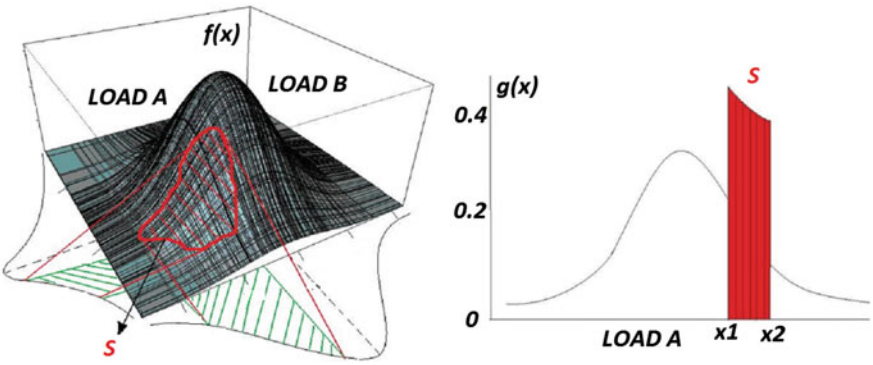

Fig. 9 Importance sampling: upward scaling of boundary region probability

and inter-dependencies. Again setting $p = 0.75$, 75 % of the points is expected from N-dimensional boundary region S, as the probability distribution is altered to produce more samples from S. Figure 9 depicts the probability reorientation by importance sampling process towards the boundary region in a 2-dimensional state space. The parameter p serves as a sliding parameter that controls the extent of biasing between a completely operational study with $p = 1$ to investment planning study with $p = 0$.

5 Case Study

The proposed sampling approach is applied for a decision tree based security assessment study for deriving operating rules against voltage stability issues on SEO region (*Système Eléctrique Ouest*, West France, Brittany), a voltage security-limited region of the French EHV system containing 5,331 buses with 432 generators supplying 83,782 MW. Figure 10 shows 400 kV network of the French system, where it can be seen that the Brittany region (highlighted in pink) is pretty weakly interconnected. During winter periods, when demand peaks, the system gets close to voltage collapse limits. Moreover the local production capabilities being far lower than the local consumption, it puts the EHV grid under pressure as the needed power comes from remote location, eventually leading to cascading phenomenon at the sub voltage levels. The busbar fault at 225 kV Cordemais bus is the most credible contingency in the Brittany region during winter period.

So in order to avoid the risk of collapse situations under such contingency events, the operator may have to resort to expensive preventive measures such as starting up close yet expensive production units. It is therefore very important to assess the risks of a network situation correctly considering uncertainties in operating conditions and obtain operating rules built with decision trees, that aid to take right decision at right time.

Section 5.1 describes the study specifications in terms of historical data used in this study, the sampling parameters and assumptions, and tools and methods used to perform power system assessments. Section 5.2 provides the numerical illustration, presenting the systematic application of stages 1 and 2 of the efficient sampling approach in Sects. 5.2.1 and 5.2.2 respectively, and finally discussing the results from the proposed method and their significance in Sect. 5.2.3 in terms of operating rule's classification accuracy and economic benefits.

5.1 Study Specifications

Data preparation: The historical database of French EHV power grid system for the study is extracted from records made every 15 s on the network by SCADA. The load in the SEO region starts to increase at the end of October, as the winter comes closer, and decreases in February. The heavily loaded period is the winter, during

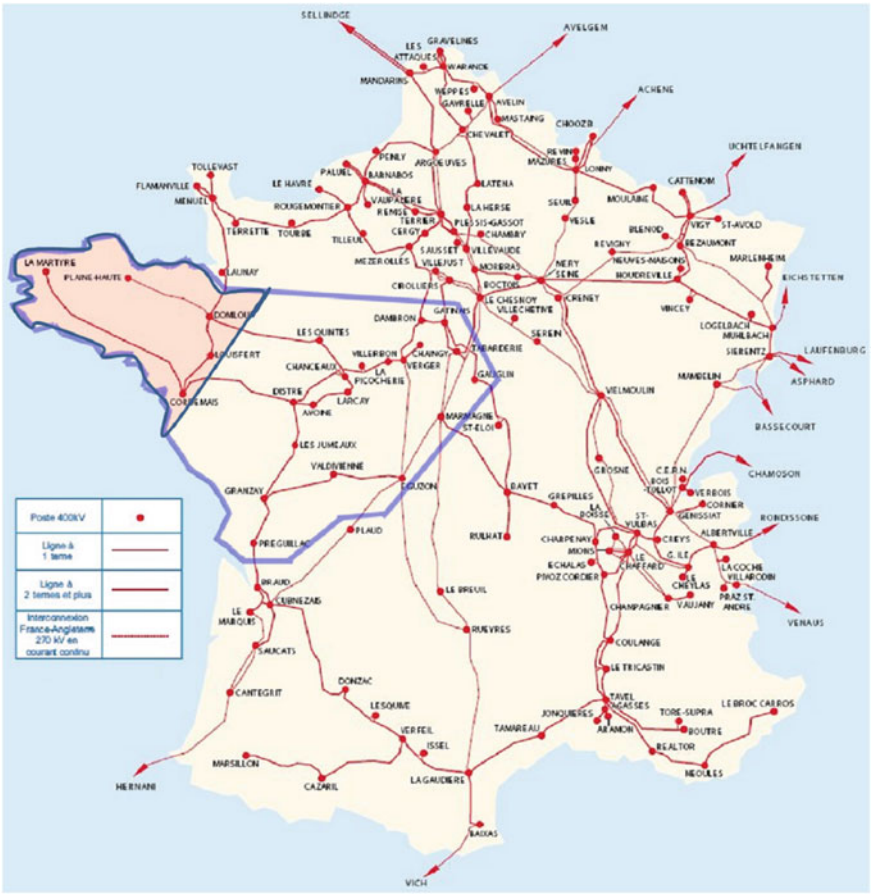

Fig. 10 French 400 kV network with SEO and Brittany highlighted

December, January, and February months. A lot of loads were shed in the month of January under stressful conditions motivated by economic and reliability considerations for system operation. The loading pattern over the year changes depending upon various factors such as, if it is winter or summer, week or week-end, day or night, peak-hours or off peak hours etc. Typically, the load is heavier during the daytime of weekdays in winter, as shown by the statistics in Table 1. Therefore, these heavily loaded periods are the most constraining in terms of voltage, and the study focuses on them for generating samples of operating conditions. Therefore, MCS is not performed on the entire year distribution, but only on those relevant periods that impact the considered stability problem.

Sampling: The pre-contingency operating conditions are generated from a base case, by considering random changes of key parameters. The basecase of SEO network considered corresponds to 2006/2007 winter, with the variable part of the

Table 1 Historical load data statistics in MW—year 2007

Period	Mean	Median	Maximum
Full year	7,729	7,640	13,607
Summer (June–September)	6,609	6,600	9,182
Winter (October–March)	8,585	8,539	13,607
Winter (December–February)	9,290	9,307	13,607
Winter (December–February)—weekdays	9,758	9,823	13,607
Winter (December–February)—week 8 h to 22 h	10,350	10,284	13,607

total baseload amounting to about 13,500 MW. The most constraining contingency is a busbar fault in the Brittany area that trips nearby generation units, which may lead to a voltage collapse under extreme conditions. The parameters sampled to generate operating conditions are variable part of total SEO load, SVC unavailability and generator group unavailability in Brittany area. The unavailability of main production units, consisting of nuclear groups at Civaux, Blayais, St-Laurent, Flamanville, and Chinon are sampled such that each of these 5 unavailabilities is represented in 1/6th of the total basecases. The unavailabilities of two SVCs at Plaine-Haute and Poteau-Rouge are sampled such that 25 % of the cases have both, 25 % do not have both and 50 % have only one of them. The variable part of total load, a continuous multivariate parameter, is sampled using our proposed efficient sampling method. The power factor of loads is kept constant. All the load samples are systematically combined with SVC and generator group unavailabilities respecting their respective sampling laws to form various operating conditions.

Contingency analysis and database generation: For each condition, an optimal power flow is performed, minimizing the production cost under voltage, current, flow constraints in N. Then consequences of busbar fault are studied with a quasi steady state simulation (QSSS) tool, where the simulation is run for 1,500 s and the contingency is applied at 900 s. Scenarios are characterized as unacceptable if any of SEO EHV bus voltage falls below 0.8 p.u or the simulation does not converge. Then a learning dataset is formed using pre-contingency attributes of every scenario (sampled at 890 s of QSSS) that drives voltage stability phenomenon, such as voltages, active/reactive power flows, productions etc., and their respective classifications. Then security rules are produced using decision tree to detect a probable voltage collapse situation contingent upon the severe event. An independent test set is used to validate the tree.

The software tools used in the study are:

1. ASSESS—Special platform for statistical and probabilistic analyses of power networks (Available at: http://www.rte-france.com/htm/an/activites/assess.jsp)
2. TROPIC—Optimal Power Flow tool, embedded with ASSESS, to create initial base cases
3. ASTRE—Simulating slow dynamic phenomena (QSSS), embedded with ASSESS
4. SAS—Statistical analysis and database processing
5. ORANGE, WEKA—Decision tree tools

5.2 Numerical Illustration

One of the major significances of this case study, apart from the demonstration of efficient training database generation for decision trees, is the consideration of system load with non-parametric multivariate distribution including the mutual correlation or inter-load dependencies. The multivariate load distribution is comprised of 640 load buses, on which the two-stage efficient sampling process is performed to generate influential operating conditions for preparing training database.

5.2.1 Stage-I: Fast Boundary Region Identification

There are 24 combinations of discrete parameters as shown in Table 2.

For the first combination, with no component unavailability, initial basecases are formed based on the sampled k homothetic stress directions using LHS. Then CPF

Table 2 Boundary identification under component combinations

S. No	SVC cases	Generator cases	P_L^{SEO}min (MW)	P_L^{SEO}max (MW)
1	None	None	11,627	12,700
2	None	Blayais	11,507	12,580
3	None	Chinon	11,474	12,547
4	None	Civaux	11,515	12,529
5	None	Flamanville	11,476	12,506
6	None	St-Laurent	11,490	12,562
7	Plaine-Haute	None	11,618	12,691
8	Plaine-Haute	Blayais	11,498	12,571
9	Plaine-Haute	Chinon	11,465	12,538
10	Plaine-Haute	Civaux	11,506	12,520
11	Plaine-Haute	Flamanville	11,467	12,497
12	Plaine-Haute	St-Laurent	11,481	12,553
13	Plaine-Rouge	None	11,608	12,681
14	Plaine-Rouge	Blayais	11,488	12,561
15	Plaine-Rouge	Chinon	11,455	12,528
16	Plaine-Rouge	Civaux	11,496	12,510
17	Plaine-Rouge	Flamanville	11,457	12,487
18	Plaine-Rouge	St-Laurent	11,471	12,543
19	Both	None	11,599	12,672
20	Both	Blayais	11,479	12,552
21	Both	Chinon	11,446	12,519
22	Both	Civaux	11,487	12,501
23	Both	Flamanville	11,448	12,478
24	Both	St-Laurent	11,462	12,534
	Boundary		**11,446**	**12,700**

Table 3 Incremental estimation of k

k	P_L^{SEO}min (MW)	P_L^{SEO}max (MW)	Gap (MW)
5	12,500	12,700	200
8	11,627	12,500	873
12	12,000	12,700	700
15	11,627	12,700	1,073
20	11,627	12,650	1,023
25	11,627	12,700	1,073

is performed to characterize the load state space and find the boundary limits of total SEO load $\{P_L^{SEO}\text{min}, P_L^{SEO}\text{max}\}$, which is found to be {11,627, 12,700} MW as shown in Table 2. The margin sensitivities are also computed along every k stress directions, which are used to estimate the change in boundary limits due to the influence of component combination change. Table 2 shows the estimated boundary limits for all the remaining combinations. The final boundary limits are estimated as 11,446 and 12,700 MW.

Table 3 shows the process of estimating k for LHS in an incremental fashion. Beyond $k = 15$, the boundary region is identified fairly consistently. The Expectation-Maximization algorithm based clustering method, when applied to historical record of stress directions, optimally grouped the stress directions into 21 clusters. This information is useful to quickly zero in on the ideal value for k.

Figure 11 shows the boundary characterization from a simulation performed for 24,000 operating conditions with randomly selected combinations of discrete

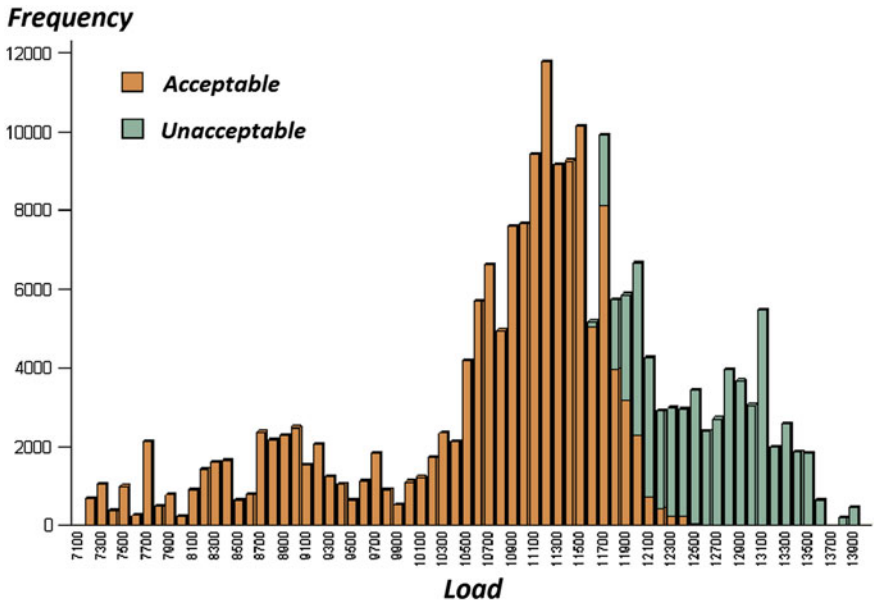

Fig. 11 Boundary characterization in total SEO load state space

parameters and loads. The boundary region (where both acceptable and unacceptable performances occur) begins approximately at around 11,500–11,700 MW and ends at around 12,500–12,700 MW. Therefore, these simulation results verify the ability of the proposed stage-I method to estimate the boundary region in the multi-dimensional operating parameter state space at a highly reduced computing requirements (i.e., about 20 CPF computations, compared to 24,000 simulations for Fig. 11).

5.2.2 Stage-II: Importance Sampling

Many MCS studies in the past have assumed a multivariate normal distribution of load data (Wan et al. 2000). But in this study, importance sampling is performed on actual empirical non-parametric distribution obtained from the projected historical data of loads. Figure 12 shows three marginal load distributions among the 640 load vectors that make up the multivariate historical data. It is seen that the multivariate distribution is made up of marginal distributions that are not exactly normal, but by visual inspection some looks close to normal, some uniform, some discrete and so on. So a multivariate Normality assumption will give misleading results.

Furthermore, these marginal distributions are not independent to model them separately as a group of normal, uniform and discrete distributions respectively; but they are mutually correlated, and the sampling method must preserve their inter-dependencies or correlations while sampling. So considering both the non-parametric nature of the marginal distributions and their mutual correlations, the whole sampling task becomes very challenging. Therefore, as mentioned in the Sect. 4.2.2, copulas are used that could efficiently work with multiple non-parametric marginal distributions and their mutual correlation (rank correlation) to produce correlated multivariate random vectors from original multivariate distribution defined by empirical historical data.

After identifying the boundary region limits, the empirical multivariate distribution of boundary region $f_1(x)$ is begotten from historical data by filtering the records within the identified boundary limits. When $p = 1$ in Eq. (8), we have complete sampling bias towards the boundary region $f_1(x)$. The inter-dependencies between various individual loads are captured in the sampling process that use

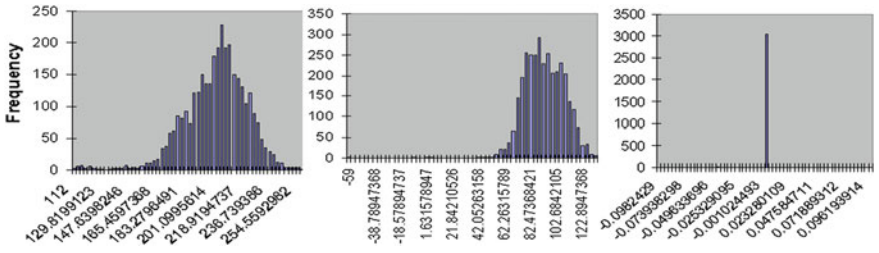

Fig. 12 Some sample marginal distributions from historical load data

copulas, and correlated multivariate random vectors from $f_1(x)$ are generated. The generated samples are for real power values only, and the reactive power at the corresponding individual load buses are obtained by maintaining the power factor constant.

5.2.3 Results and Discussions

The training database generated within the boundary region contains 2,852 operating conditions. The test database includes 1,976 independent instances unseen by training set, covering a wide range of operating conditions. The candidate attributes available for rule formation consists of 46 and 102 node voltages at 400 and 225 kV voltage levels respectively, reactive power flows (*Q flow*) in 16 tie lines, real power reserve (*P res*) in SEO from 10 generator group's, and reactive power reserve (*Q res*) in SEO from 10 generator group's and 2 SVCs.

Table 4 shows the effectiveness of various combinations of attribute sets in terms of classification accuracy and error rates. Accuracy is defined as the percentage of points correctly classified, false alarm rate is defined as the ratio of total misclassified unacceptable instances among all unacceptable classifications, and risk rate is defined as the ratio of total misclassified acceptable instances among all acceptable classifications. The attribute set "400 kV + Q res" proves to be a good set with lowest risk and high classification accuracy. It has to be noted that the accuracy listed in the Table 4 are for trees that are pruned by restricting the minimum number of instances per leaf node.

Effect of Bias Factor "P"

This section sheds light on the quantitative impact of biasing the sampling process towards the boundary region by presenting results for various values of bias factor

Table 4 Attribute set selection

Attribute Set	Accuracy	False alarm	Risk	Tree size
400 kV + Q res	87.9079	0.193	0.073	15
Q res	87.7159	0.183	0.083	15
225 kV	82.8215	0.243	0.124	15
400 kV + 225 kV	82.7255	0.253	0.12	15
400 kV + 225 kV + Q res	82.6296	0.236	0.132	13
All	82.6296	0.236	0.132	13
225 kV + Q res	82.4376	0.231	0.139	13
400 kV	80.8061	0.231	0.166	17
Q flow	75.5278	0.325	0.191	23
P res	73.8004	0.402	0.169	13

"p". Specifically, two aspects are discussed, namely (a) computational requirements and accuracy, and (b) economic benefits.

(a) *Computation, Accuracy and Tree Size*: Fig. 13a–d show the total SEO load probability distribution from sampled operating conditions as the sliding factor p increases from the base value in $f(x)$ to 1 (bias towards *boundary*).

Table 5 shows the results when validated using the test database, which confirms that as the sampling of operating conditions is biased towards the boundary region, the entropy of the database increases (a quantitative indicator of information content) and even with lesser database size higher accuracy for decision tree is obtained. The error rates, namely *false alarms* and *risk*s are both simultaneously reduced to a great degree. It was also found that as the sampling is biased more towards the boundary region, the size of the decision tree required for good classification also decreased. This is due to the ability of database to capture high information content (i.e., the variability of performance measure across the security boundary) even with smaller number of instances.

(b) *Economically beneficial rules*: Table 6 presents the influence of efficient sampling in producing economical rules. The table shows that for the various possibilities of the decision tree top node attribute among the most influential attributes, the database generated from within boundary region with $p = 1$ finds rules with attribute thresholds that are always less conservative than from the database that was generated with $p = 0$, i.e., from entire operational state space.

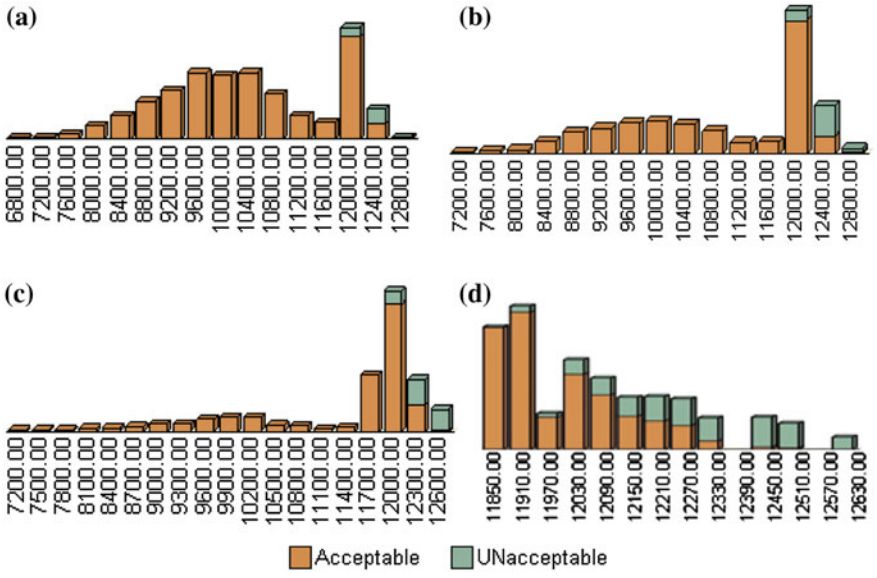

Fig. 13 Effect of p on sampled total SEO load probability distribution. **a** $p = 0.25$. **b** $p = 0.50$. **c** $p = 0.75$. **d** $p = 1.0$

Table 5 Performance based on sampling bias

p	Size	Entropy	Accuracy	False alarm	Risk
Base	17,748	0.7423	92.51	0.063	0.091
0.25	13,840	0.7716	93.4211	0.064	0.068
0.50	9,932	0.8181	94.9899	0.049	0.051
0.75	6,025	0.9038	96.0526	0.038	0.041
1.0	2,852	0.9993	97.5202	0.021	0.03

Figure 14 shows operational rule formed using two attributes, namely reactive reserves at Chevire unit and Chinon unit respectively. The operating conditions shown in the Fig. 14 are from the entire database. It can be noticed that the rules formed using the database exclusively from the boundary region is providing more operating conditions to be exploited in real time situations, than the rule derived using the database from entire region; because of the increased knowledge and clarity of the boundary limits.

Sampling Strategies Comparison

Table 7 shows the comparison results of two different sampling approaches, namely,

1. Importance sampling (IS) of boundary region, with load distribution modeled with multivariate normal (MVN) distribution (pruned tree).
2. Importance sampling of boundary region, with load distribution modeled with correlated non-parametric multivariate distribution (MVD) (pruned tree).
3. Same as case 2, with un-pruned tree.

Table 6 Economic benefit from efficient sampling

Top Node	*p = 0*	*p = 1*
Cordemais voltage	401.64 kV	399.88 kV
Domloup voltage	397.56 kV	394.51 kV
Louisfert voltage	399.1 kV	396.46 kV
Plaine-Haute voltage	392.26 kV	387.21 kV
Chevire unit reactive reserve	131.38 Mvar	90.76 Mvar
Chinon unit reactive reserve	1,127.54 Mvar	694.62 Mvar
Cordemais unit reactive reserve	70.97 Mvar	16.23 Mvar
Total SEO region reactive reserve	7,395.88 Mvar	6,510.36 Mvar
Plaine-Haute SVC output	11.82 Mvar	13.64 Mvar
Poteau-Rouge SVC output	16.3 Mvar	22.03 Mvar

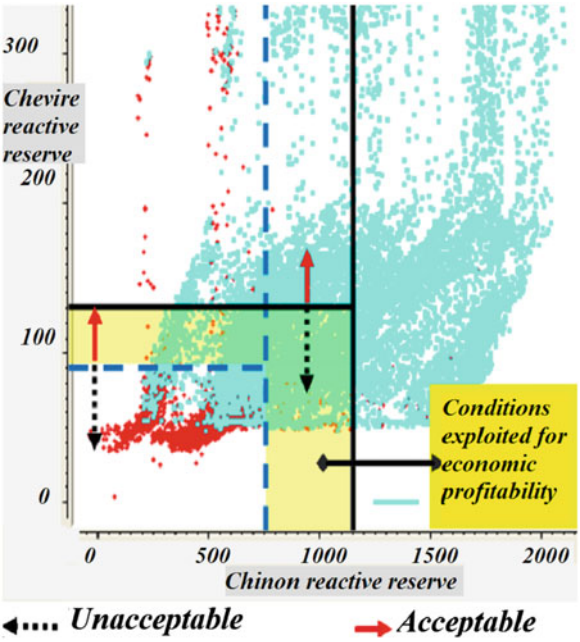

Fig. 14 Economical benefit of operational rules from efficient sampling

Table 7 Comparison between different sampling strategies

Sampling strategy	Size	Accuracy	False alarm	Risk
IS (MVN—pruned)	2,879	80.6142	0.142	0.228
IS (MVD—pruned)	2,852	87.0951	0.094	0.178
IS (MVD)	2,852	97.5202	0.021	0.03

It can be seen from Table 7 that the database produced by importance sampling of correlated-MVD state space definitely shows better performance. When the trees are pruned for operator's convenience of usage the accuracy decreases, which can be improved using the right-hand side loop as shown in the Fig. 3. It also performs better than sampling from MVN load space, which is conventional assumption in many studies due to trivial modeling requirements.

The significance of sampling from correlated-MVD, i.e., capturing the inter-load dependencies, than from MVN is even strongly vindicated by Fig. 15 that shows the top 5 critical attribute locations produced by decision trees from respective databases. The contingency event is shown by a red star. The location of 5 critical monitoring attributes as well as their sequence in the tree matters. Compared to MVN, all the 5 top locations found by correlated-MVD sampling strategy are very interesting ones, with the top node being reactive reserve at a big nuclear plant Chinon, the node in the next level of the tree is closer to the contingency location, the next nodes (3 and 4) in the tree deals with the two SVC locations in Brittany, and finally the attribute of node 5 is right at the contingency location.

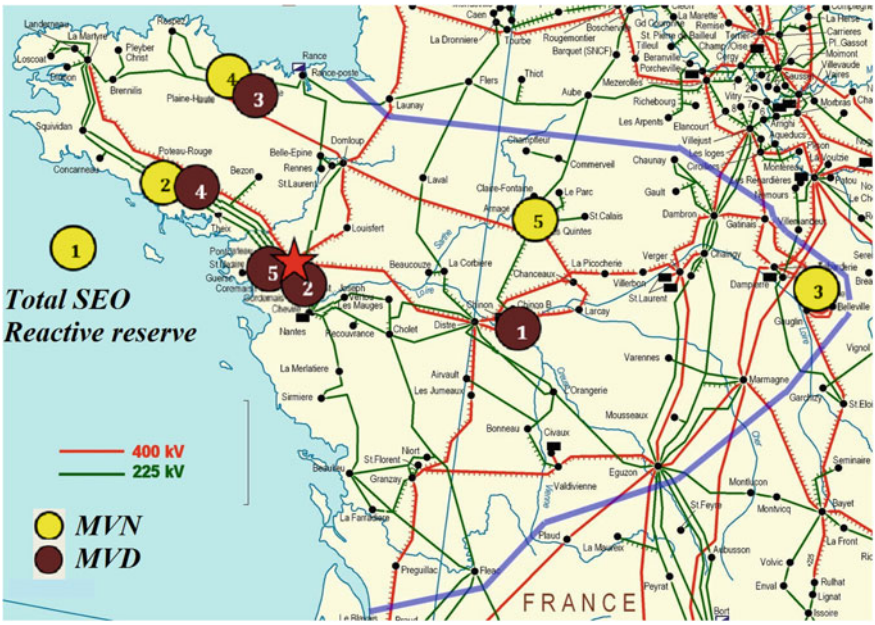

Fig. 15 Critical monitoring locations from decision tree: MVD versus MVN

6 Conclusion

The proposed efficient sampling method based on importance sampling idea is one of the first to be used in power systems for making decision tree based learning methods effective. The thrust of the proposed sampling procedure was to re-orient the sampling process to focus more heavily on points for which post-contingency performance is close to the threshold, i.e., boundary region that contains operating conditions influential for rule formation. The primary goal was to increase the information content in the learning database while reducing the computing requirements, and consequently obtain operational rules that are more accurate for usage in real-time situations.

The developed efficient training database approach was applied for deriving operational rules in a decision tree based voltage stability assessment study on RTE-France's power grid. The results showed that the generated training database enhances rules' accuracy at lesser computation compared to other traditional sampling approaches, when validated on an independent test set. The chapter also emphasized the significance of sampling from non-parametric correlated-multivariate load distribution obtained from historical data, as it is more realistic. Doing so also ensures generating operating rules that provide higher classification accuracy and economics, and selecting interesting monitoring locations that are closer to the contingency event, as corroborated by the results. In order to reduce

the computational burden in characterizing multivariate load state space, a linear sensitivity based method supported by Latin hypercube sampling of homothetic stress directions was developed for quickly characterizing the multivariate load state space for various combinations of component unavailabilities. This aided in identifying the boundary region with respect to post-contingency performance measure quickly.

The future directions of research include:

- *Application for other stability problems*: The efficient database generation approach can also be applied to other stability problems such as rotor angle stability, out of step etc. In these problems the performance measure's trajectory sensitivities will have to be used to reduce the computational cost in identifying the boundary region.
- *Optimal placement of Phasor Measurement Units* (*PMUs*): The high information content in the training database generated from the proposed efficient sampling method enables finding the most important system attributes for power system's security state monitoring. This concept is highly beneficial in finding the optimal placement of PMUs and extracting relevant knowledge from those PMUs for advancing data-driven power system operation and control.
- *Application in the reliability assessment of Special Protection System* (*SPS*): The main difference between deriving operating rules and SPS logic are:
  - The SPS logic is automated.
  - The SPS logic is not only limited to critical operating condition detection with respect to some stability criteria, but also involves automatic corrective action to safeguard the system against impending instability.

Even though many works exist that correspond to SPS "process level" design procedures and failure assessments, there are important questions to be answered about SPS operations from a 'system view-point', such as:

- Are there system operating conditions (topology, loading, flows, dispatch, and voltage levels) that may generate a failure mode for the SPS?
- Are there two or more SPS that may interact to produce a failure mode?

So the objective of this research will be to develop a decision support tool to perform SPS failure mode identification, risk assessment and logic re-design from a 'systems view'. The efficient scenario processing method presented in this chapter has tremendous scope to be used in biasing the sampling process such that SPS failure modes (including multiple SPS interactions) can be identified, risk levels may be estimated, and accordingly the logic may be re-designed using the efficient decision tree process.

Acknowledgments The author acknowledges Professor James D. McCalley at Iowa State University (Ames, Iowa, USA), Sebastien Henry at RTE-France (Versailles, France), and Samir Issad at RTE-France (Versailles, France) for their valuable support during the course of this research project.

References

Abed, A. M. (1999). WSCC voltage stability criteria, under voltage load shedding strategy, and reactive power reserve monitoring methodology. In IEEE Power Engineering Society Summer Meeting (pp. 191–197), July 18–22, 1999, Edmonton, Alta. doi:10.1109/PESS.1999.784345.

Beshir, M. J. (1999). Probabilistic based transmission planning and operation criteria development for the Western Systems Coordinating Council. In IEEE Power Engineering Society Summer Meeting (pp. 134–139), July 18–22, 1999, Edmonton, Alta. doi:10.1109/PESS.1999.784334.

Billinton, R., & Li, W. (1994). *Reliability assessment of electric power systems using Monte Carlo methods*. New York: Plenum Press.

Billinton, R., Salvaderi, L., McCalley, J. D., Chao, H., Seitz, Th, Allan, R. N., et al. (1997). Reliability issues in today's electric power utility environment. *IEEE Transactions Power Systems, 12*(4), 1708–1714.

Cholley, P., Lebrevelec, C., Vitet, S., & De Pasquale, M. (1998). Constructing operating rules to avoid voltage collapse: A statistical approach. In *Proceedings of International Conference on Power System Technology*, *POWERCON '98* (pp. 1468–1472), August 18–21, 1998, Beijing. doi:10.1109/ICPST.1998.729331.

Chowdhury, A. A., & Koval, D. O. (2006). Probabilistic assessment of transmission system reliability performance. In IEEE Power Engineering Society General Meeting, Montreal, Que. doi:10.1109/PES.2006.1709096.

Cutsem, T. V., Wehenkel, L., Pavella, M., Heilbronn, B., & Goubin, M. (1993). Decision tree approaches to voltage security assessment. In *IEE Proceedings on Generation, Transmission and Distribution* (Vol. 140, No. 3, pp. 189–198).

Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer.

Dobson, I., & Lu, L. (2002). New methods for computing a closest saddle node bifurcation and worst case load power margin for voltage collapse. *IEEE Transactions Power Systems, 8*(3), 905–913.

Dy-Liacco, T. E. (1997). Enhancing power system security control. *IEEE Computer Applications in Power, 10*(3), 38–41.

Foody, G. M. (1999). The significance of border training patterns in classification by a feed forward neural network using back propagation learning. *International Journal of Remote Sensing, 20*(18), 3549–3562.

Genc, I., Diao, R., Vittal, V., Kolluri, S., & Mandal, S. (2010). Decision tree-based preventive and corrective control applications for dynamic security enhancement in power systems. *IEEE Transactions Power Systems, 25*(3), 1611–1619.

Gentle, J. E. (1998). *Random number generation and Monte Carlo methods*. Newyork: Springer.

Greene, S., Dobson, I., & Alvarado, F. L. (1997). Sensitivity of the loading margin to voltage collapse with respect to arbitrary parameters. *IEEE Transactions on Power Systems, 12*(1), 262–272.

Henry, S., Lebrevelec, C., and Schlumberger, Y. (1999). Defining operating rules against voltage collapse using a statistical approach: The EDF experience. In *International Conference on Electric Power Engineering*, *PowerTech Budapest*, August 29–September 2, 1999, Budapest, Hungary. doi:10.1109/PTC.1999.826461.

Henry, S., Bréda-Séyès, E., Lefebvre, H., Sermanson, V., & Béna, M. (2006). Probabilistic study of the collapse modes of an area of the French network. In *Proceedings of the 9th International Conference on Probabilistic Methods Applied to Power Systems* (pp. 1–6), June 11–15, 2006, Stockholm. doi:10.1109/PMAPS.2006.360261.

Henry, S., Pompee, J., Bulot, M., and Bell, K. (2004a). Applications of statistical assessment of power system security under uncertainty. In *International Conference on Probabilistic Methods Applied to Power Systems* (pp. 914–919). September 16–16, 2004, Ames, IA.

Henry, S., Pompee, J., Devatine, L., Bulot, M., and Bell, K. (2004b). New trends for the assessment of power system security under uncertainty. In *IEEE PES Power Systems Conference and Exposition* (pp. 1380–1385), Oct 10–13, 2004. doi:10.1109/PSCE.2004.1397731.

Hormann, W., Leydold, J., & Derflinger, G. (2004). *Automatic non-uniform random variate generation*. Newyork: Springer.

Jacquemart, Y., Wehenkel, L., & Pruvot, P. (1996). Practical contribution of a statistical methodology to voltage security criteria determination. In *Proceedings of the 12th Power Systems Computation Conference* (pp. 903–910).

Jiantao, X., Mingyi, H., Yuying, W., & Yan, F. (2003). A fast training algorithm for support vector machine via boundary sample selection. In *Proceedings of the International Conference on Neural Networks and Signal Processing* (pp. 20–22), December 14–17, 2003, Nanjing. doi:10.1109/ICNNSP.2003.1279203.

Krishnan, V., Liu, H., and McCalley, J. D. (2009). Coordinated reactive power planning against power system voltage instability. In *IEEE/PES Power Systems Conference and Exposition* (pp. 1–8), March 15–18, 2009, Seattle, WA. doi:10.1109/PSCE.2009.4839926.

Lebrevelec, C., Schlumberger, Y., & De Pasquale, M. (1999). An application of a risk based methodology for defining security rules against voltage collapse. In IEEE Power Engineering Society Summer Meeting (pp. 185–190), Jul 18–22, 1999, Edmonton, Alta. doi:10.1109/PESS.1999.784344.

Lebrevelec, C., Cholley, P., Quenet, J.F., & Wehenkel, L. (1998). A statistical analysis of the impact on security of a protection scheme on the French power system. In *Proceedings of International Conference on Power System Technology*, *POWERCON* (pp. 1102–1106), Aug 18–21, 1998, Beijing. doi:10.1109/ICPST.1998.729256.

Li, W., & Choudhury, P. (2007). Probabilistic transmission planning. *IEEE Power and Energy Magazine, 5*(5), 46–53.

Long, B., & Ajjarpau, V. (1999). The sparse formulation of ISPS and its application to voltage stability margin sensitivity and estimation. *IEEE Transactions on Power Systems, 14*(3), 944–951.

Martigne, H., Cholley, P., King, D., & Christon, J. (2001). Statistical method to determine operating rules in the event of generator dropout on EDF French Guyana Grid. In *Proceedings of IEEE Power Tech* (pp. 1–5), Sep 10–13, 2001, Porto. doi:10.1109/PTC.2001.964599.

Niimura, T., Ko, H. S., Xu, H., Moshref, A., and Morison, K. (2004). Machine learning approach to power system dynamic security analysis. *IEEE PES Power Systems Conference and Exposition* (pp. 1084–1088), October 10–13, 2004. doi:10.1109/PSCE.2004.1397549.

Papaefthymiou, G., & Kurowicka, D. (2009). Using copulas for modeling stochastic dependence in power system uncertainty analysis. *IEEE Transactions Power Systems, 24*(1), 40–49.

Paul J., Bell K. (2004). A flexible and comprehensive approach to the assessment of large-scale power system security under uncertainty. *International Journal of Electrical Power & Energy Systems*, *26*(4), 265–272.

Pierre, J., Lebrevelec, C., & Wehenkel, L. (1999). Automatic learning methods applied to dynamic security assessment of power systems. In *International Conference on Electric Power Engineering. PowerTech Budapest*, Aug 29–Sept 2, 1999, Budapest, Hungary. doi:10.1109/PTC.1999.826612.

Rencher, A. (1995). *Methods of multivariate analysis*. New York: Wiley.

Ripley, B. D. (1987). *Stochastic Simulation*. New York: Wiley.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo method*. New York: Wiley.

Schlumberger, Y., Lebrevelec, C., & De Pasquale, M. (1999). Power systems security analysis-new approaches used at EDF. IEEE Power Engineering Society Summer Meeting (pp. 147–151), Jul 18–22, 1999, Edmonton, Alta. doi:10.1109/PESS.1999.784337.

Schlumberger, Y., Pompee, J., & De Pasquale, M. (2002). Updating operating rules against voltage collapse using new probabilistic techniques. In *IEEE/PES Transmission and Distribution Conference and Exhibition*: *Asia Pacific* (pp. 1139-1144), October 6–10, 2002. doi:10.1109/TDC.2002.1177638.

Senroy, N., Heydt, G. T., & Vittal, V. (2006). Decision tree assisted controlled islanding. *IEEE Transactions Power Systems, 21*(4), 1790–1797.

Singh, C., & Mitra, J. (1997). Composite system reliability evaluation using state space pruning. *IEEE Transactions Power Systems, 12*(1), 471–479.

Thisted, R. A. (1988). *Elements of statistical computing*. New York: Chapman and Hal Ltd.

Unger, E. A., Harn, L., and Kumar, V. (1990). Entropy as a measure of database information. In *Proceedings of the Sixth Annual in Computer Security Applications Conference* (pp. 80–87), December 3–7, 1990, Tucson, AZ. doi:10.1109/CSAC.1990.143755.

Wan, H., McCalley, J. D., & Vittal, V. (2000). Risk based voltage security assessment. *IEEE Transactions Power Systems, 15*(4), 1247–1254.

WECC. (2003, April). NERC/WECC planning strandards. Available at, http://www.wecc.biz/documents/library/procedures/planning/WECC-NERC_Planning%20Standards_4-10-03.pdf.

Wehenkel, L. (1997). Machine learning approaches to power-system security assessment. *IEEE Expert, IEEE Intelligent Systems and Their Applications, 12*(5), 60–72.

Wehenkel, L. (1998). *Automatic learning techniques in power systems*. Berlin: Kluwer Academic Publishers.

Wehenkel, L., Glavic, M., Geurts, P., & Ernst, D. (2006). Automatic learning of sequential decision strategies for dynamic security assessment and control. In *IEEE Power Engineering Society General Meeting*, *Montreal*, *Que*. doi: 10.1109/PES.2006.1708874.

Witten, I. H., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with Java implementations*. San Fransisco, CA: Morgan Kaufmann Publishers.

Wyss, W. G., & Jorgensen, K. H. (1998). A user's guide to LHS: Sandia's latin hypercube sampling software. *Sandia National Laboratories Report SAND98-0210*, Albuquerque, NM.

Xiao, F., & McCalley, J. D. (2007). Risk-based security and economy tradeoff analysis for real-time operation. *IEEE Transactions Power Systems, 22*(4), 2287–2288.

Yu, X., & Singh, C. (2004). Expected power loss calculation including protection failures using importance sampling and SOM. In *IEEE Power Engineering Society General Meeting* (pp 206–211), June 6–10, 2004. doi:10.1109/PES.2004.1372787.

Zhou, G., & McCalley, J. D. (1999). Composite security boundary visualization. *IEEE Transactions Power Systems, 14*(2), 725–731.

Zhou, Q., Davidson, J., & Fouad, A. A. (1994). Application of artificial neural networks in power system security and vulnerability assessment. *IEEE Transactions Power Systems, 9*(1), 525–532.