Using Vine Copulas to Generate Representative System States for Machine Learning

Konstantelos, Ioannis; Sun, Mingyang; Tindemans, Simon; Issad, Samir; Panciatici, Patrick; Strbac, Goran

DOI
10.1109/TPWRS.2018.2859367

Publication date
2019

Document Version
Final published version

Published in
IEEE Transactions on Power Systems

Using Vine Copulas to Generate Representative System States for Machine Learning

Ioannis Konstantelos, *Member, IEEE*, Mingyang Sun, *Member, IEEE*, Simon H. Tindemans, *Member, IEEE*, Samir Issad, *Member, IEEE*, Patrick Panciatici, *Senior Member, IEEE*, and Goran Strbac, *Member, IEEE*

***Abstract*—The increasing uncertainty that surrounds electricity system operation renders security assessment a highly challenging task; the range of possible operating states expands, rendering traditional approaches based on heuristic practices and ad hoc analysis obsolete. In turn, machine learning can be used to construct surrogate models approximating the system's security boundary in the region of operation. To this end, past system history can be useful for generating anticipated system states suitable for training. However, inferring the underlying data model, to allow high-density sampling, is problematic due to the large number of variables, their complex marginal probability distributions and the nonlinear dependence structure they exhibit. In this paper, we adopt the C-Vine pair-copula decomposition scheme; clustering and principal component transformation stages are introduced, followed by a truncation to the pairwise dependency modeling, enabling efficient fitting and sampling of large datasets. Using measurements from the French grid, we show that a machine learning training database sampled from the proposed method can produce binary security classifiers with superior predictive capability compared to other approaches.**

***Index Terms*—Copulas, data clustering, machine learning, Monte Carlo simulation, parametric statistics, principal component analysis, risk assessment, stochastic dependence, uncertainty analysis.**

NOMENCLATURE

A. Data Sets and Distributions

Distributions and random vectors are denoted in bold script.

$\boldsymbol{Z}$	Random variable drawn from the 'true' joint distribution we want to approximate.
$\boldsymbol{Z}^e$	Random variable defined by the observations data set $\mathbb{Z}$.
$\hat{\boldsymbol{Z}}$	Random variable that approximates $\boldsymbol{Z}$ i.e., the output of the statistical model fitted to $\mathbb{Z}$.
$\mathbb{Z}$	Data set of historical observations.
$\hat{\mathbb{Z}}$	Sampled data set.
$z^{(t)}$	A single observation of $\mathbb{Z}$ at some time $t \in \{1, \ldots, N\}$.
$F^e_{\mathbb{Z}i}$	Empirical distribution of variable i in data set $\mathbb{Z}$.

B. Scalars and Model Parameters

n	Number of variables in $\mathbb{Z}$.
N	Number of observations in $\mathbb{Z}$.
K	Number of clusters.
N^c	Number of observations of $\mathbb{Z}$ belonging to cluster c.
m	Number of variables modelled using C-Vine copulas.
$\hat{\phi}_{k,j}$	Family of bivariate copula coupling variables $\mathcal{U}_{k\|j-1:1}$ and $\mathcal{U}_{j\|j-1:1}$.
$\hat{\theta}_{k,j}$	Parameters of bivariate copula coupling variables $\mathcal{U}_{k\|j-1:1}$ and $\mathcal{U}_{j\|j-1:1}$.
N_s	Number of realizations to be sampled from $\hat{\boldsymbol{Z}}$.

I. INTRODUCTION

THE large-scale integration of intermittent generation sources has led to a significant increase in the uncertainty characterizing the operation of electricity systems [1]. This is exacerbated by the growing interconnection between independently-operated markets and the advent of shiftable load elements, such as electric vehicles. Existing security tools relying on heuristic rules and ad hoc analysis close to real-time are quickly becoming obsolete, in view of the expanding operation range and complexity of electricity systems. Another major limitation of existing approaches is that they do not present a scalable way for expanding the list of contingencies analyzed. As such, they usually focus upon examining faults that are highly probable or have occurred in the past. This way the system operator remains blind to previously-unseen contingencies which may turn out to be critical.

At the same time, instrumentation is increasing at all system levels with the deployment of Phasor Measurement Units (PMUs) and cross-exchange of information between operators across borders and voltage levels [2]. Large-scale data collection presents immense untapped potential for supporting the secure operation of the grid. Data from the system's past history can be combined with high-fidelity simulations and be used to train binary classifiers for determining whether a unseen operating point is safe or unsafe after a specific contingency [3]. The use of such classifiers foregoes the need to carry out simulations near real-time that suffer from computational constraints. As

Manuscript received November 14, 2017; revised May 15, 2018; accepted July 22, 2018. Date of publication July 24, 2018; date of current version December 19, 2018. Paper no.TPWRS-01712-2017. *(Corresponding author: Mingyang Sun.)*

I. Konstantelos, M. Sun, and G. Strbac are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW72AZ, U.K. (e-mail: i.konstantelos@imperial.ac.uk; mingyang.sun11@imperial.ac.uk; g.Strbac@imperial.ac.uk).

S.-H. Tindemans is with the Department of Electrical Sustainable Energy, Delft University of Technology, Delft 2628, The Netherlands (e-mail: s.h.tindemans@tudelft.nl).

S. Issad is with the RTE/DMA, Versailles 78005, France (e-mail: samir.issad@rte-france.com).

P. Panciatici is with the RTE-France, Versailles 78005, France (e-mail: patrick.panciatici@rte-france.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2018.2859367

such, data-driven security monitoring presents a scalable way to accommodate the increasing uncertainties and an expanding set of contingencies. In this paper we study the topic of efficient database generation for constructing good quality predictors.

In the past, many authors have investigated the construction of surrogate models of the security inference problem using machine learning techniques. In [3] and [4] the authors introduced the concept of building proxies to the transient stability problem using machine learning. The suitability of different training features was assessed and decision trees were shown to be capable of accurate predictions. Authors in [5] and [6] use decision trees to configure preventive control schemes. The training database for the dynamic security assessment is constructed by sampling uncertain variables using independent Gaussian distributions fitted to historical data and then iteratively refining the sample space using importance sampling principles. Authors in [7] focus on identifying dynamic stability from PMU data using decision trees; many different loading conditions are sampled while the loading factor of each bus is assumed constant. Tests carried out on a 68-bus system achieved high prediction accuracy. The author in [8] examined different classifiers such as random trees and support vector machines.

In general, the predictive capability of classifiers is highly dependent upon the data used for training. Machine learning algorithms require a database that has a good representation of all class values so that new instances can be accurately classified. For the security assessment problem, this translates to having a training database that (i) has high diversity and coverage (ii) includes both secure and unsecure post-fault operation (iii) is in close vicinity to the instances to be classified in the future, thus exhibiting good generalization capability. Note that the training database refers to pre-fault operating points and is independent of the list of contingencies to be simulated (i.e., computing the post-fault operating points) which can include faults that have never occurred in the past. Historical data constitute a natural starting point for constructing a training dataset. Thereafter, different approaches can be taken to enhance the already-available information. For example, authors in [9] and [10] use an entropy-based importance sampling method to identify informative stress directions for generating load scenarios.

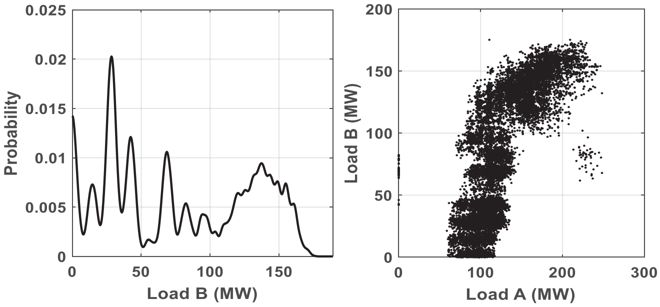
In this paper we focus on fitting a statistical model to the historical data and sampling at high densities. There are several advantages to inferring a probabilistic model instead of direct use of past measurements (empirical distribution). Most importantly, a model provides the ability to generate training databases of arbitrarily large size. This is necessary for generating training data sets of high variability for machine learning tasks. Moreover, probabilistic models can readily be combined with other techniques, such as importance sampling, as proposed in [9]. A final aspect is the dramatic reduction of memory requirements which can be a problem with data-intensive applications. Given that future transmission systems will be equipped with thousands of sensors and millions of smart meters, the ability to compress large data volumes in concise models without substantial information loss is highly desired.

In the past there have been several efforts towards building statistical models of high-dimensional stochastic variables. Copula models, in particular, have become increasingly popular in many fields of application from financial market modeling [11] to weather-related research [12] and consumer profiling [13]. Copulas have also been used in the past to capture the relation between different stochastic attributes in the context of electrical energy system operation [14] and planning [15]. Authors in [16] and [17] use vine copulas to capture the dependency between the outputs of different wind farms. The same principle is applied in [18] to electric vehicle usage data to generate loading scenarios. Other applications include multi-attribute modeling, as in [19], which focuses on the effect of wind power on hydro-dominated systems. Another approach to multivariate modelling is the use of graphical models such as Bayesian networks (BNs). BNs can be hampered by a number of issues that limit their modeling capability, such as the inability to control the marginal distributions and the inability to capture complex dependence structures. Authors in [20] show that BNs can be viewed as a specific case of vine copulas, meaning that BNs entail an inherently constrained structure while vine copulas, by definition, are more flexible in modelling complex data sets.

Beyond fitting statistical models to a data set, it is also possible to adopt a model-free approach. In particular, the use of Generative Adversarial Networks (GANs) has been proposed with great success in the field of computer vision [21]. In the area of power systems, researchers in [22] have recently used GANs to sample wind farm and solar plan output time series. GANs have been developed very recently and there are numerous open questions regarding their interpretability, hyper-parameterization, convergence [23], loss function definition and capability to actually learn the underlying distribution [24]. For these reasons, in this paper we focus primarily on statistical modelling techniques.

Despite the numerous applications, all existing approaches deal with only a handful of variables; there has been little effort to develop a unified framework for modeling high-dimensional dependent stochastic variables. Recently, the authors in [25] investigated the potential for combining data clustering, dimension reduction and vine copula techniques for high-dimensional stochastic modeling along with examining different validation methods. In this research we extend this work by giving specific modeling details and investigating the suitability of this model as a database generation tool for machine learning tasks in highly stochastic multivariate electricity systems. Finally, note that the main differences of this paper compared to [25] are: (i) showcase of the full model formulation as well as extensive discussion on critical modelling choices (ii) propose model truncation using a multivariate Gaussian copula which results in severe model size reduction (iii) comprehensive case study demonstrating the importance of generating high densities of representative system states for data-driven security assessment (iv) computationally-efficient algorithms for model parameterization and sampling; note the full source code has been made open source in [40].

The work presented here has been implemented in iTesla [26], a platform for security analysis in very large grids that uses machine learning to infer security rules. Given the huge computational requirements of analyzing the security of national-level systems [27], efficient database generation is of paramount importance.



(a) (b)

Fig. 1. (a) Marginal probability distributions of 5-minute load measurements over 3 months from a bus in the region of Nancy, France (March 2012). (b) Non-linear dependence between load measurements of two other buses in the same region.

The remaining paper is organized as follows. Section III describes the challenges to be overcome. Sections IV and V present the different components of the proposed workflow. Section VI presents a case study showing the superior performance of the proposed method compared to existing approaches for generating training databases for the security assessment problem. Section VII summarizes and concludes.

II. Modelling Challenges

Let the set of historical observations consists of N observations of n interdependent injections and loads $z_i^{(t)}$, where $i \in \{1, \ldots, n\}$ is the nodal index and $t \in \{1, \ldots, N\}$ the observation index, usually associated with a snapshot time. We will denote each snapshot as a vector $\vec{z}^{(t)} \in \mathbb{R}^n$, and the set of all observations as

$$\mathcal{Z} = \left\{ \vec{z}^{(t)} | t \in \{1, \ldots, N\} \right\}. \tag{1}$$

It is instructive to consider the random variable $\boldsymbol{Z}^e$ that is defined by the empirical observations $\mathcal{Z}$. The empirical samples are considered to be drawn from an unknown 'true' distribution, represented by the continuous random variable $\boldsymbol{Z}$. The objective of our work is the construction of an approximate random variable $\hat{\boldsymbol{Z}}$ that approximates $\boldsymbol{Z}$, by fitting it to the set of observations $\mathcal{Z}$. Subsequently we compare the use of $\hat{\boldsymbol{Z}}$ and $\boldsymbol{Z}^e$ for machine learning applications.

First of all, one of the main challenges of the task at hand is that almost all variables of interest, such as active/reactive loads and renewables injections have highly non-standard distributions, complicating the use of purely parametric statistical models. For reference, a typical marginal distribution of load levels is shown in Fig. 1(a).

Secondly, the stochastic variables of interest exhibit non-linear dependence rendering all traditional statistical methods that assume independence or solely rely on Pearson's correlation coefficient inadequate. For example, a scatter plot displaying the dependence between two load buses is shown in Fig. 1(b). The plot suggests asymmetric dependence, which is beyond the capabilities of linear models.

Beyond the statistical properties of the variables, another aspect that renders the modeling task challenging is the dataset size. The electricity system of a medium-sized country can involve several thousand stochastic attributes including active and reactive load levels, power injections due to intermittent sources and uncontrollable cross-border flows. Such large multivariate models can quickly encounter practical limitations due to exponentially-increasing computation times for model parameterization. In this research we use Principal Component Analysis (PCA) to project the high-dimensional stochastic signal onto an 'information-ordered' space so as to focus the subsequent complex modeling tasks on a reduced subset of variables, rendering the proposed approach tractable.

The final challenge is the large number of historical observations. For each stochastic variable, there can be thousands of recorded measurements available, rendering the process of identifying a single parametric model that fits the data a very challenging task. To tackle this issue, appropriate techniques can be used to partition the observations into groups so as to differentiate between system modes that result in signals with radically different statistical behavior. This paper proposes an enhanced approach by considering a mixture of copula vines using k-means clustering.

III. Variable Truncation C-Vine Copula

Given the highly complex dependence between stochastic variables in electricity systems, a suitable model must capture not only linear correlations but also higher-order statistics such as tail-to-tail dependencies. Copula-based methods are explicitly suitable for modeling such dependencies. We present here a particular implementation of a copula-based model, the *Variable Truncation C-Vine method*, which is adapted to modelling complex dependencies in high-dimensional systems. It consists of the following steps: (i) cluster the data (ii) apply Spearman's Principal Component Analysis to identify the most important dimensions for detailed dependency modelling; (iii) use of the C-Vine Pair Copula Construction to model the dependencies between the m most important dimensions; (iv) use of a multivariate Gaussian copula to model the dependencies with and between the remaining $n - m$ dimensions.

A. Model Clustering

In general, networks operate across a range of qualitatively different operating regimes. For example, electrical consumption patterns during peak hours are determined by fundamentally different drivers when compared to nighttime consumption. To address this issue, historical observations can be assumed to not come from a single underlying model, but from multiple distinct models. K-means clustering is used to partition N observations into K clusters, where each observation belongs to the cluster with its nearest mean as defined below:

$$\left\{\vec{\mu}^1, \ldots, \vec{\mu}^K\right\} = \underset{\left\{\vec{\mu}^1, \ldots, \vec{\mu}^K\right\}}{\operatorname{argmin}} \left\{ \sum_{c=1}^{k} \sum_{\vec{z} \in \mathcal{Z}^c} \left\| \vec{z} - \vec{\mu^c} \right\|^2 \right\} \tag{2}$$

where $\mathcal{Z}^c$ is the set of points that is closer to point $\vec{\mu}^c$ than to any other $\vec{\mu}^{c\prime \neq c}$ (using a Euclidean metric). This operation partitions the original population $\mathcal{Z}$ into k disjoint sets $\{\mathcal{Z}^c\}_{c=1}^K$

with a variable number of elements $N^c = |\mathcal{Z}^c|$. The optimal choice for the number of clusters is data-dependent; the larger the number of clusters, the simpler the structure of each cluster becomes, but the model is more susceptible to overfitting. As discussed in [28], there are various methods to determine the optimal number of clusters. Note that hereafter we drop the explicit dependence on the cluster index c, with the implicit understanding that models are constructed independently for each cluster.

B. Spearman's Principal Component Analysis

The variable truncation C-vine model that is described below has a hierarchical structure; the dependencies between dimensions with a lower index are modelled with fewer transformations than those with higher indices. Using this structure to our advantage, we place high-variance modes early in the cascade, followed by those with lower variance. PCA can be used to perform this variable ordering and reduce the computational complexity of the proposed method. However, PCA is highly sensitive to the skewness and magnitude of variables and as such, it is good practice to transform data prior to PCA. In general, the choice of transformation type is highly data-dependent. In this case, the Probability Integral Transform (PIT) is applied to each data cluster $\mathcal{Z}$ to construct $\mathcal{Y} = (Y_1, .., Y_n)$, where each variable Y_i has been transformed through its respective empirical cumulative distribution function (ecdf) F^e_{Zi} as in (4). Applying PCA to PIT-transformed data is known as Spearman's PCA and is equivalent to performing PCA in terms of the Spearman's correlation matrix [29]. After extensive testing, it was found to perform better than conventional PCA on the French power system data set, where variables follow non-Gaussian marginal distributions and are highly dissimilar.

$$\mathcal{Y}_i = F^e_{Z_i}(\mathcal{Z}_i), \text{ for } i \in \{1, \ldots, n\} \tag{3}$$

Here and throughout the paper, whenever a mapping is applied to a set, it implies application to each element, i.e., every $z_i^{(t)} \in \mathcal{Z}_i$. The $n \times n$ covariance matrix of $\mathcal{Y}$ can be factorized as:

$$\text{cov}(\mathcal{Y}) = \Psi \Lambda \Psi^T \tag{4}$$

The matrix Ψ^T defines an orthogonal mapping onto the principal component (PC) domain, in which dimensions are sorted in order of decreasing variance. If the variables are significantly correlated, it is possible to focus the modelling effort on a small number m of principal components (i.e., $m, \ll n$) while losing only a fraction of the full attribute information. We transform the original data from cluster c into its principal component space as follows

$$\mathcal{X} = \left\{\vec{x} | \vec{x} = \Psi^T \vec{y}, \vec{y} \in \mathcal{Y}\right\} \tag{5}$$

From here on we use x to denote coordinates in the PC space.

C. Copulas and the C-Vine Construction

The basic concept of copulas follows from Sklar's theorem [41]. Consider n random variables $X_1, X_2, \ldots, X_n$ with marginal density functions $f_i(x_i)$ and associated distribution functions $F_i(x_i)$. The joint density function is given by

$$f(x_{1:n}) = c_{12\ldots n}(F_1(x_1), \ldots, F_n(x_n)) \times f_1(x_1) \ldots f_n(x_n) \tag{6}$$

where we use the notation $x_{1:n}$ as a shorthand for the sequence $x_1, x_2, \ldots, x_n$. The *copula density* $c_{12\ldots n} : [0,1]^n \rightarrow \mathbb{R}$ describes the dependence between uniform random variables $\{U_1, \ldots, U_n\} = \{F_1(X_1), \ldots, F_n(X_n)\}$. The copula representation provides a convenient way to separate the marginal distributions of X_i from their dependency structure. The copula connecting all variables is unique if all marginal distribution functions are continuous; for independent variables $c_{12\ldots n} = 1$. The practical problem of interest is the identification of the best-fitting parametric copula to the transformed empirical data $\mathcal{X}$, represented by the random vector $\boldsymbol{X}$.

For the bivariate case, there is a well-investigated and rich variety of copula families [30]. However, in the case of higher dimensions, the dependency patterns that may exist between large numbers of variables are far more complex than for the bivariate case. Attempting to capture this dependency structure using a single multivariate parametric copula can be restrictive and lack the flexibility required. The pair-copula construction (PCC) approach represents a way of constructing high-dimensional complex models of multivariate dependence, by extending the bivariate theory to an arbitrary number of dimensions [31]. It was first introduced by Joe in [30] and developed in more detail in [32], [33] and [34]. The main idea of PCC is to decompose a multivariate distribution into a product of bivariate copulas by using recursive conditioning. A joint probability density is recursively factorized as follows:

$$\begin{aligned} f(x_{1:n}) = f_1(x_1) \cdot f_{2|1}(x_2|x_1) \cdot f_{3|21}(x_3|x_2, x_1) \cdot \ldots \\ \cdot f_{m|m-1:1}(x_m|x_{m-1:1}) \\ \cdot f_{n:m+1|m:1}(x_{n:m+1}|x_{m:1}). \end{aligned} \tag{7}$$

Each of the univariate conditional probability distributions can be decomposed into (conditional) pairwise copulas using

$$\begin{aligned} f_{j+i|j:1}(x_{j+i}|x_{j:1}) &= \frac{f_{j+i,j|j-1:1}(x_{j+i}, x_j|x_{j-1:1})}{f_{j|j-1:1}(x_j|x_{j-1:1})} \\ &= c_{j+i,j|j-1:1}\left(F_{j+i|j-1:1}, F_{j|j-1:1}\right) f_{j+i|j-1:1}(x_{j+i}|x_{j-1:1}) \end{aligned} \tag{8}$$

where the second equality follows by expanding the bivariate function $f_{j+i,j|j-1:1}(x_{j+i}, x_j|x_{j-1:1})$ using (7)- an operation that is not affected by conditioning on $x_{j-1:1}$. For clarity we have used the shorthand notation $F_{j+i|j-1:1}$ to represent the conditional distribution function $F_{j+i|j-1:1}(x_{j+i}|x_{j-1:1})$. This equation is used recursively to express all conditional densities for dimensions $1, \ldots, m$ in (7) as products of bivariate copula densities and the corresponding marginal densities, resulting in the decomposition

$$f\left(x_{1:n}\right)=\prod_{j=1}^{m-1}\prod_{i=1}^{m-j}c_{j+i,j|j-1:1}\left(F_{j+i|j-1:1}\ ,F_{j|j-1:1}\right)$$
$$\cdot\left(\prod_{k=1}^{m}f_k\left(x_k\right)\right)\cdot f_{n:m+1|m:1}\left(x_{n:m+1}|x_{m:1}\right). \tag{9}$$

Note that the distribution in dimensions $m+1,\ldots,n$ is left unspecified for now. Through this decomposition it becomes possible to retain the separation between marginals and dependency modeling, while permitting the use of arbitrary bivariate copula families, capturing wide range of different dependency structures. It is important to remark that the order of pairwise conditioning in (7) and (8) is not unique. In fact, there are $2m!$ ways to decompose an m-dimensional joint probability function [33]. The specific recursive conditioning structure defined above is known as the Canonical Vine (C-Vine) 0. Although in theory all $2m!$ decompositions are valid and equivalent, in practical applications parametric fits are performed and the order does matter. This is because any inaccuracy due to an ill-fitting copula can propagate 'downstream', affecting parameter choice of all recursively-defined copulas. The C-Vine structure is well-suited to our application because PCA assigns a crude importance ordering to the variables $X_1,\ldots,X_n$.

For notational clarity, we define the conditioned random variables $U_{j+i|j:1}\equiv F_{j+i|j:1}(X_{j+i}|X_{j:1})$. Each is a uniform variable on the unit domain, and, by construction, the fully conditioned unit variables $V_j\equiv U_{j|j-1:1}$ are mutually independent. Furthermore, we will make use of the h-function notation introduced in [33] to denote conditional distribution functions

$$F_{j+i|j:1}\left(x_{j+i}|x_{j:1}\right)=\int_{-\infty}^{x_{j+i}}\frac{f_{j+i,j|j-1:1}\left(x',x_j|x_{j-1:1}\right)}{f_{j|j-1:1}\left(x_j|x_{j-1:1}\right)}dx'$$
$$=h_{j+i,j}\left(F_{j+i|j-1:1}(x_{j+i}|x_{j-1:1}),F_{j|j-1:1}(x_j|x_{j-1:1})\right) \tag{10}$$

with

$$h_{j+i,j}\left(u,v\right)\equiv\int_0^u c_{j+i,j|j-1:1}\left(w,v\right)\mathrm{d}w \tag{11}$$

The h-functions therefore define the recursive relation between the conditioned unit random variables as

$$U_{j+i|j:1}=h_{j+i,j}\left(U_{j+i|j-1:1},U_{j|j-1:1}\right). \tag{12}$$

Expressions for the h-functions of common copula functions are available in [33]. The process of fitting a C-Vine model to the data is summarized in Algorithm 1. Data is successively conditioned on variables $1,\ldots m-1$, by fitting bivariate copulas $\hat{c}_{k,j|j-1:1}$ and transforming observations using the corresponding h-functions. The fitting of a single copula function consists of determining an appropriate copula family and corresponding parameters; the latter are determined via the Maximum Likelihood Estimation (MLE) method [33]. There are different criteria that can be used to select the best-fitting family such as the Vuong test, goodness-of-fit (GOF) test, the Akaike information criterion (AIC), and the Bayesian inference criterion (BIC) 0. Authors in [11] showed AIC to be the best-performing criterion and as such we adopt it in this research. The selected copula family and parameters associated with the transformation $\hat{h}_{j+i,j}$ are denoted by $\hat{\phi}_{j+i,j}$ and $\hat{\theta}_{j+i,j}$, respectively.

Algorithm 1: Construction of Variable Truncation C-Vine.

```
for i ← 1, ..., n
    $\mathcal{U}_i = F^e_{X_i}(\mathcal{X}_i)$ (transform all observations to unit domain)
for j ← 1, ..., m − 1
    for $k$ ← j + 1, ..., m
        $\hat{\phi}_{k,j},\ \hat{\theta}_{k,j}$ ← bivariate copula fit on
        $(\mathcal{U}_{k|j-1:1},\mathcal{U}_{j|j-1:1})$
        $\mathcal{U}_{k|j:1} = \hat{h}(\mathcal{U}_{k|j-1:1},\mathcal{U}_{j|j-1:1};\hat{\phi}_{k,j},\ \hat{\theta}_{k,j})$
for i ← 1, ..., m
    $\mathcal{W}_i = \mathcal{U}_{i|i-1:1}$ (select $m$ conditioned variables)
for i ← m + 1, ..., n
    $\mathcal{W}_i = \mathcal{U}_i$ (augment with $n-m$ unconditioned variables)
$\hat{R}$ ← $n-$dimensional Gaussian copula fit on $\mathcal{W}$
```

D. Truncation With Gaussian Copula

The C-Vine structure is used to model the dependencies between the first m dimensions in principal component space. Although it is a very flexible modelling technique, it is computationally intensive, because it requires the fitting of $m(m-1)/2$ bivariate copulas, each involving the evaluation of d copula families. For this reason we truncate the C-Vine construction at $m\ll n$ and use a less accurate but more computationally efficient Gaussian multivariate copula to model the dependencies in remaining dimensions, represented by the conditional PDF $f_{n:m+1|m:1}(x_{n:m+1}|x_{m:1})$ in Eq. (9).

Let us consider the coordinate transformation $X\rightarrow U\rightarrow V$ implied by the C-Vine for the first m dimensions. Because this mapping is one-to-one, we may replace the conditioning on $x_{m:1}$ by an equivalent conditioning on $v_{m:1}$. This change enables us to express the conditional PDF using an n-dimensional copula as follows, invoking Sklar's theorem and the mutual independence of V_i:

$$f_{n:m+1|m:1}\left(x_{n:m+1}|x_{m:1}\right)=\frac{f'_{n:m+1,m:1}\left(x_{n:m+1},v_{m:1}\right)}{f'_{m:1}\left(v_{m:1}\right)}=$$
$$c'_{n:m+1,m:1}\left(F_n,\ldots,F_{m+1},F_{m|m-1:1},\ldots,F_1\right)\times\prod_{i=m+1}^{n}f_i\left(x_i\right) \tag{13}$$

We find that the dependency structure is described by a single copula that relates the dependent unit variables $U_{m+1},\ldots,U_n$ and the mutually independent unit variables $V_1,\ldots,V_n$.

A simple approximation would be to assume independence, which equates to setting $c'_{n:m+1|m:1}=1$, so that the PDF of $X_{n:m+1}$ is a simple product of the marginal distributions. Numerical experiments show that this requires the inclusion of a significant number of dimensions m in the C-Vine to achieve a good fit (result not shown). Instead we propose to explicitly take

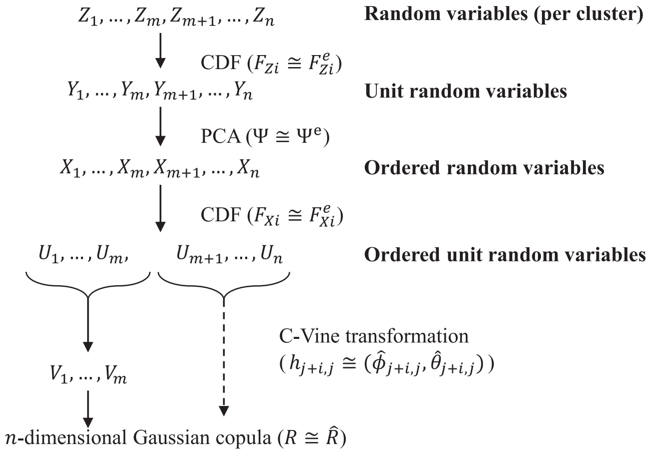

$Z_1, \dots, Z_m, Z_{m+1}, \dots, Z_n$
Random variables (per cluster)
CDF ($F_{Zi} \cong F_{Zi}^e$)
$Y_1, \dots, Y_m, Y_{m+1}, \dots, Y_n$
Unit random variables
PCA ($\Psi \cong \Psi^e$)
$X_1, \dots, X_m, X_{m+1}, \dots, X_n$
Ordered random variables
CDF ($F_{Xi} \cong F_{Xi}^e$)
$U_1, \dots, U_m,$
$U_{m+1}, \dots, U_n$
Ordered unit random variables
C-Vine transformation
($h_{j+i,j} \cong (\hat{\phi}_{j+i,j}, \hat{\theta}_{j+i,j})$)
$V_1, \dots, V_m$
n-dimensional Gaussian copula ($R \cong \hat{R}$)

Fig. 2. Variable truncation C-Vine procedure (for a single cluster). The successive transformations are indicated by both their 'true' and 'approximate' or empirical versions.

linear correlations into account by approximating $c'_{n:m+1|m:1}$ with a single n-dimensional Gaussian copula, parameterized by the correlation matrix $\hat{R}$.

This choice is included in Algorithm 1, as a final step of parameter estimation. The full model construction process (transformations and parameters) is illustrated in Fig. 2. We note that the proposed Variable Truncation C-Vine method differs from the 'Truncated C-Vine' described in [11], which conditions all n features on the m first variables, instead of only the first m. In the case where $m \gg n$, this requires ≈2n/m times the number of copula evaluations.

IV. Model Sampling Algorithm

The model consists of a sequence of steps: k-means clustering, PCA, C-Vine copula construction and multivariate Gaussian copula fitting. Stochastic random variables can be generated by traversing the modules in reverse order. Approaches on how to simulate from vine copulas were first proposed in [20] and presented in detail in [33]. A computationally efficient algorithm for sampling from the Variable Truncation C-Vine model is shown in Algorithm 2.

First, a random vector $\overrightarrow{w} \in [0,1]^n$ is generated according to the multivariate Gaussian distribution parameterized by $\hat{R}$. The first m components are the starting point for the C-Vine sampling procedure. The first coordinate, $\hat{v}_1$, which in a C-Vine is the 'governing' variable, is considered independent to all other variables. The other coordinates are transformed to the dependent unit coordinates $\hat{u}_i$ by recursive un-conditioning using $\hat{h}^{-1}$ (inverse of $\hat{h}$ with respect to the first parameter). The resulting unit coordinates $\hat{u}_{1:m}$ are combined with the remaining coordinates $\hat{u}_{m+1:n} = \hat{w}_{m+1:n}$ and they are transformed back to the principal component domain through the inverse empirical distribution functions $\{F_{Xi}^{-1}\}_{i=1}^n$. The transformation of the resulting vector by the matrix Ψ results in a single random realization $\{\hat{y}_i\}_{i=1}^n$. Each dimension must be subsequently transformed through the inverse empirical distribution function F_{Zi}^{-1}.

Algorithm 2: Truncated C-Vine Sampling Algorithm (for a Single Sample).

```
Sample $\{\hat{w}_i\}_{i=1}^n \sim$ multivariate Gaussian $(\hat{R})$.
$\hat{v}_{1:m} = \hat{w}_{1:m}$
$\hat{u}_1 = \hat{v}_1$
for k $\leftarrow 2, \dots,$ m
    $\hat{u}_k = \hat{v}_k$
    for $j \leftarrow$ k $- 1, \dots, 1$
        $\hat{u}_k = \hat{h}^{-1}(\hat{u}_k, \hat{v}_j; \hat{\phi}_{k,j}, \hat{\theta}_{k,j})$
$\hat{u}_{m+1:n} = \hat{w}_{m+1:n}$
for i $\leftarrow 1, \dots,$ n
    $\hat{x}_i = (F_{X_i}^e)^{-1}(\hat{u}_i)$
$\overrightarrow{\hat{y}} \leftarrow \Psi \overrightarrow{\hat{x}}$, truncated to $[0,1]^n$
for i $\leftarrow 1, \dots,$ n
    $\hat{z}_i = (F_{Z_i}^e)^{-1}(\hat{y}_i)$
```

Algorithm 2 presents the process for generating a single sample. In practice, the user will want to generate a large number of samples N_s. This requires sampling all K C-Vine cluster-models. Note that each cluster model has an associated probability π^c which depends on N^c, the number of historical observations grouped in cluster c i.e., $\pi^c = N^c/N$. As such, stratified sampling can be used, where each cluster-model is sampled to generate $N_s^c = N_s \pi^c$ samples. If N_s is sufficiently large (as would be the case for security monitoring), then it can be rounded up or down to the nearest integer variable with negligible accuracy loss. The sampled data set $\hat{\mathbb{Z}}$ is constructed by appending the output of all K cluster models.

Note that all parameterization and sampling methods described in Sections IV and V were developed in MATLAB and the code has been made available at [40].

V. Case Study

A. *Historical and Sampled Dataset*

In this case study we use a modified version of the IEEE 118-bus system to study the use of surrogate models for identifying the security boundary. The original system consisting of 54 generators and 186 lines has been modified to also include 10 wind farms of size 100 MW each. To create a historical database $\mathbb{Z}$ of sufficient complexity, we use a dataset provided by RTE, the French system operator. The dataset contains high-voltage active load and wind generation 5-minute measurements between January and March 2012; 14,250 observations spanning over 7,000 nodes.

From this set, 118 demand buses and 10 wind generators from the area of Nancy were chosen at random and 'mapped' (i.e., scaled according to the maximum value defined in the coincident peak snapshot) to the 118-bus test system [36]. Note that the case study uses a DC power flow approximation and thus only active loads were considered in the model construction and sampling. If required, the presented method could accommodate reactive loads as additional variables in a straightforward way.

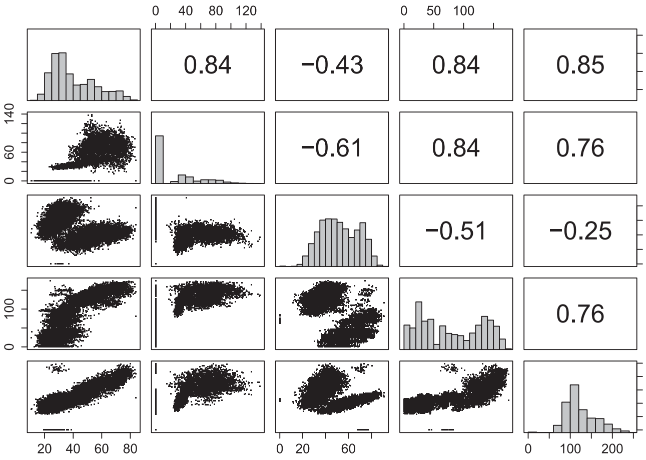


Fig. 3. Marginal and scatter plots of the historical dataset $\mathcal{Z}$.

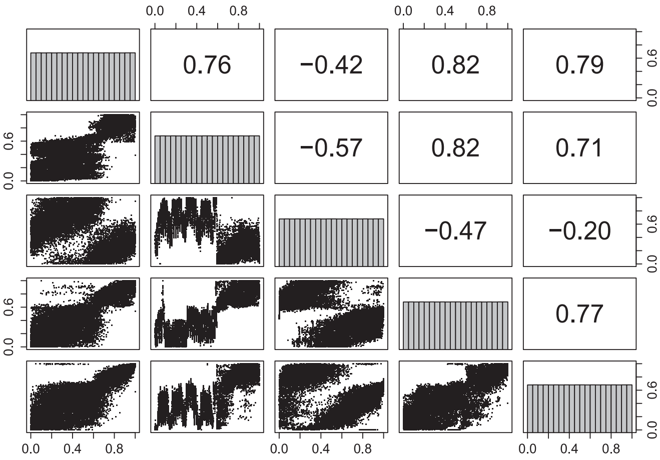


Fig. 4. Marginal and scatter plots of the dataset $\mathcal{Y}$ after applying the PIT.

A sample population of 40,000 realizations was generated using the proposed method. The number of clusters was set to 10, as determined by various clustering validity indicators, while an 97.5% variance criterion was used to determine m, the number of variables selected for C-Vine modeling. The copula families modelled were Clayton, Frank, Gaussian, Gumbel, Student-t as well as their 90°, 180° and 270° rotated versions. The algorithm was implemented in MATLAB and run on an Intel Xeon PC with 8 cores. C-Vine model parameterization and sampling was carried out in parallel for each cluster and took 34 minutes. CPU times are shown in Table I.

B. Visual Exploration of Transformation Stages

We first focus on a small example dataset where five variables were chosen at random (from $\mathcal{Z}$). Our aim is to illustrate the different transformation steps and provide justification for modelling choices involved in the proposed workflow; the small number of dimensions enables us to validate the approach through graphical exploration (note that no clustering has been used here). Fig. 3 presents the dataset by using a matrix of scatter plots where each plot shows the dependency structure between two variables. For example, the top-left scatter plot shows concordance between the first and second variable. Histograms of the five variables appear along the diagonal. The upper triangular matrix shows Pearson's correlation for each bivariate combination. As can be seen below, the marginal and bivariate distributions are highly non-Gaussian and complex. The first step is to apply the PIT on $\mathcal{Z}$. The resulting dataset, $\mathcal{Y}$, is shown in Fig. 4; all variables have uniformly distributed marginal pdf's and scatter plots are in the unit square. The second step is PCA. The resulting dataset $\mathcal{X}$ is shown in Fig. 5 where variables are ordered from highest to lowest eigenvalue.

As can be seen above, the dependence structure between the first and second PCs is clearly non-Gaussian. However, the concordance between lower-ranked PCs is increasingly more Gaussian (i.e., elliptical). The fact that the bivariate relations between lower-ranked PCs are highly Gaussian was also verified by performing the Doornick-Hansen test [39]. The same pattern holds true for the marginal pdf's; the first PC follows a bimodal distribution, while the subsequent PCs are increasingly Gaussian. The Shapiro-Wilk normality test [42] was used to verify the increasing univariate normality of lower-ranked PCs. These observations validate the modelling choices described in this paper; a series of transformation steps successfully concentrate non-Gaussian dependencies in the higher-ranked variables which are then fitted with a C-Vine model. The residual dependency structure, which is less complex, is fitted with a simpler Gaussian copula model.

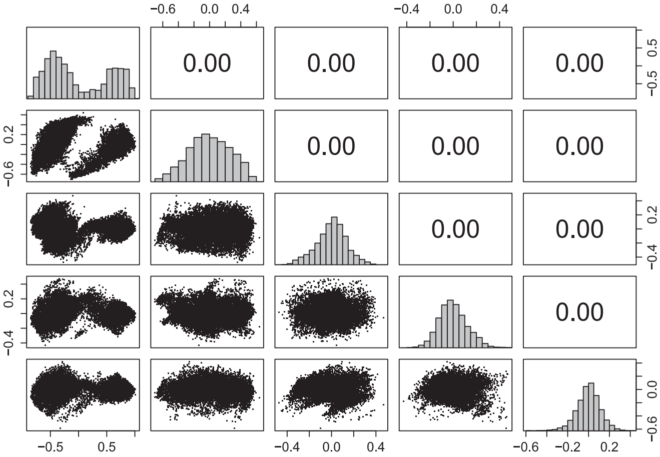


Fig. 5. Marginal and scatter plots of the dataset $\mathcal{X}$, after PCA.

C. Two-Sample Test Validation

We proceed with model validation via two-sample tests on the full 128–variable dataset $\mathcal{Z}$. Three different methods are compared; the proposed C-Vine method, Multivariate Gaussian Distribution (MGD) and Multivariate Gaussian Copula (MGC). The Kolmogorov-Smirnov (K-S) test [43] is used to examine the reconstruction of marginal distributions, while the multivariate energy test [44] is used to examine whether the original dependency structure and marginals are well captured.

The historical database was randomly split into one training and one test set that comprise 80% and 20% of the original population respectively. Three models (MGD, MGC and the

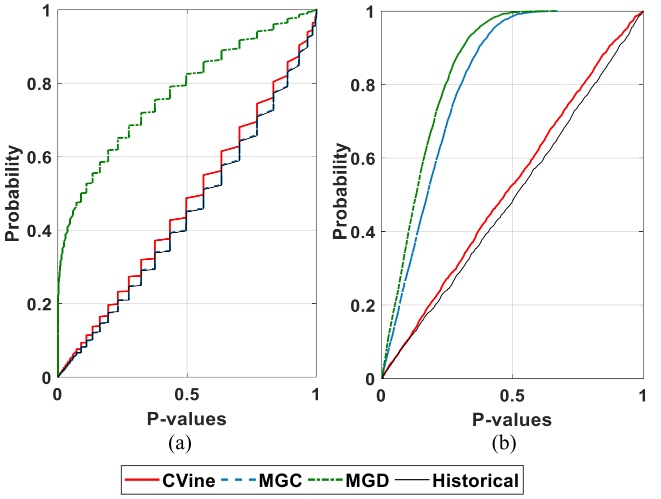


Fig. 6. Cumulative distribution of p-values. (a) For K–S. (b) For energy tests.

proposed C-Vine) were trained on this training set and 40,000 realizations were sampled from each model. Subsequently, 1,000 sets of 200 observations were drawn at random from each sampled population and compared to 200 observations randomly drawn from the historical data test set. This process resulted in a total of 1,000 energy tests and 128,000 K-S tests (one for each variable). Results of these tests are shown in Fig. 6.

For both tests, the p-values should be uniformly distributed under the null hypothesis (i.e., historical and sampled populations have been drawn from the same model). Note that this is the case when the historical database is compared against itself (i.e., we compare the train and the test partitions of the historical data); the cumulative distribution of p-values lies on the diagonal indicating a high degree of similarity of the underlying joint distributions, as expected. As can be seen above, the proposed C-Vine method outperforms both MGD and MGC methods by a large margin due to the combination of its superior ability to model marginals (shared with MGC) and its flexibility in modeling the data dependence structure.

D. Security Analysis Results

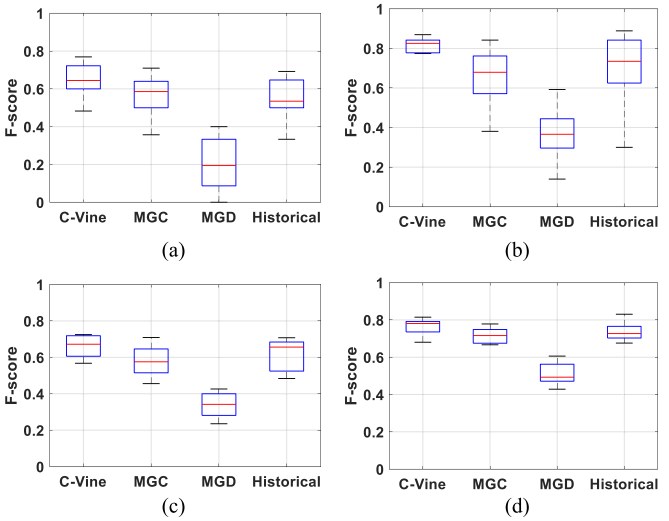
In this section we demonstrate the suitability of the proposed C-Vine method for generating training databases for machine learning a system's security boundary. We focus on a set of contingencies $\mathcal{C}$, consisting of four line outages; lines 54, 71, 148, and 154 (denoted L54 etc.) which are the most highly-utilized lines in the 118-bus system. Note that, in principle, the user is free to choose an arbitrarily large/complex set of contingencies for analysis. Our aim in this case study is to compare the performance of four training database construction approaches; the proposed C-Vine method, MGD, MGC and the historical dataset $\mathcal{Z}$ itself. A 10-fold validation scheme was used, where 10 sub-populations $\{\mathcal{Z}_k\}_{k=1}^{10}$, each containing 90% of N observations, were randomly drawn without replacement from the original dataset $\mathcal{Z}$. The remaining 10% of each $\mathcal{Z}_k$, denoted $\mathcal{Z}_k^V$, was held out as a corresponding testing set. No effort was made to reduce temporal correlations between training and test sets, because these affect all methods equally.

TABLE I
COMPUTATION TIMES FOR PARAMETERIZATION AND SAMPLING

Stage	CPU Time (s)
Data Clustering	1.95
PCA	0.02
C-Vine Parameterization	2,051.45
Gaussian Copula Parameterization	2.56
Total time for model construction	2,055.98
Time to generate a single realization	0.0214

TABLE II
AVERAGE DT TEST ERRORS FOR DIFFERENT TRAINING DATABASES

c		Historical	MGD	MGC	C-Vine
L54	NPV	99.21%	98.90%	98.84%	99.19%
	PPV	59.61%	30.02%	60.36%	65.25%
	ACC	98.27%	97.24%	97.92%	98.39%
L71	NPV	96.84%	98.27%	98.50%	98.55%
	PPV	71.54%	43.18%	68.68%	75.77%
	ACC	96.84%	94.85%	96.67%	97.13%
L148	NPV	99.19%	98.52%	99.17%	99.67%
	PPV	53.73%	22.05%	61.72%	65.98%
	ACC	99.19%	98.52%	99.17%	99.35%
L154	NPV	99.82%	99.04%	99.77%	99.86%
	PPV	69.42%	52.97%	63.76%	82.76%
	ACC	99.59%	98.70%	99.50%	99.73%
Mean	NPV	98.77%	98.68%	99.07%	99.32%
	PPV	63.58%	37.05%	63.63%	72.44%
	ACC	98.47%	97.33%	98.32%	98.65%



(a) (b) (c) (d)

Fig. 7. F-score boxplots for contingencies. (a) L148. (b) L139. (c) L54. (d) L71.

For each method, a sampled dataset $\hat{\mathcal{Z}}_k$ was generated. For each loading-wind scenario s in $\hat{\mathcal{Z}}_k$ the DC Optimal Power Flow (DC OPF) problem was solved for the optimal dispatch schedule $\boldsymbol{u}_s$. Note that, as in typical DC OPF problems, all generators were assumed to be available and their generation cost to be fully known. The use of a more sophisticated market simulation layer to increase sample diversity, such as sampling unit availability, is possible in a straightforward manner. Subsequently, each schedule was checked for post-fault feasibility under each

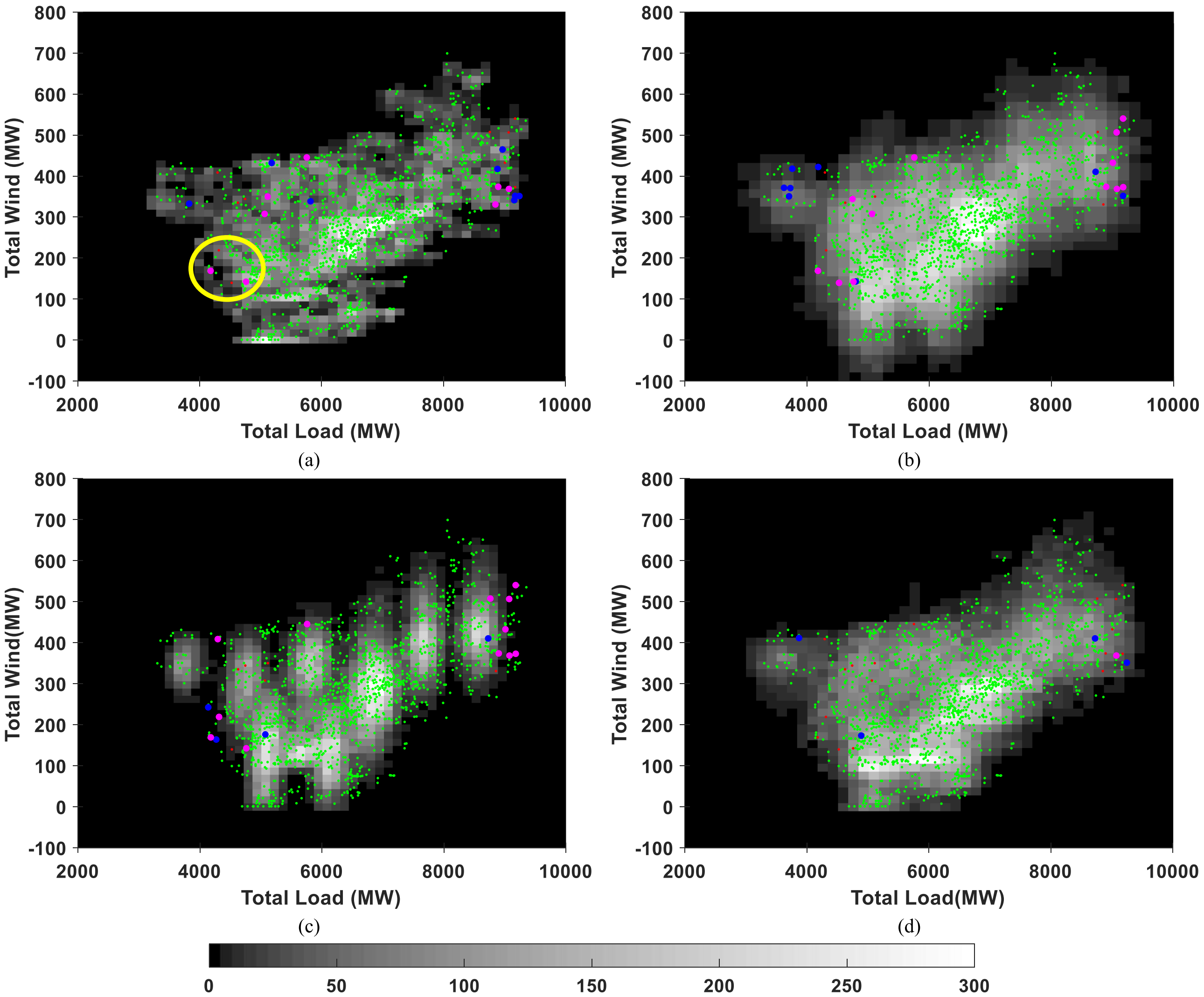


Fig. 8. Decision boundaries for DTs trained. (a) On historical data. (b) On MGD. (c) On MGC. (d) C-Vine datasets for contingency L71.

contingency c in $\mathcal{C}$; conventional generators were allowed to deviate by ±10% from their pre-fault schedule to simulate corrective response action. As such, each system post-fault state was associated with an unsafe/safe label denoting post-fault load curtailment or not respectively.

Subsequently, for each contingency c and fold k, a decision tree (DT) was trained on tuples of pre-fault bus angles and post-fault stable/unstable labels and then tested on its ability to predict the labels corresponding to the test set $\mathcal{Z}_k^V$. Matlab 2017a default parameters were used for decision tree training.

For each contingency the Negative Predictive Value (NPV), Positive Predictive Value (PPV), and Accuracy (ACC) of the constructed DTs, averaged across the 10 folds, are listed in Table II (best performers are underlined). Accuracy is defined as the ratio of correctly-classified operating points, while PPV and NPV indicate the proportions of correctly-predicted unsafe and safe scenarios respectively. In Fig. 7 we show box plots of the DTs' f-scores (harmonic mean of recall and precision) for the four contingencies analyzed.

The proposed C-Vine method outperforms both MGC, MGD and historical data in the vast majority of cases. Although the average difference in terms of accuracy and NPV is modest, the difference in terms of PPV is very substantial. Note that PPV is of particularly high interest in the context of security assessment. The same pattern is observed in the case of f-scores; the proposed method has higher mean values and lower variance under all contingencies, indicating the C-Vine model results in superiorly-trained classifiers.

In general, MGD and MGC underperform because the fitted models do not capture well the characteristics of $\mathcal{Z}$, resulting in training databases that are much less relevant for inferring the post-fault security of $\mathcal{Z}_k^V$. In contrast, the original dataset and the proposed method generate more relevant samples. This point is illustrated in Fig. 8 where we show the training domain and corresponding predictions for DTs trained on four different datasets stemming from the same cross-validation fold of $\mathcal{Z}$. Plots are shown in two dimensions; total load and total wind. A grayscale density denotes each DT's training domain; as shown in the bottom legend, white color signifies areas where about 300 samples were used in the training dataset, whereas dark areas correspond to fewer training points. Small green and red dots denote correctly-classified safe and unsafe operating points respectively. In contrast, larger-sized blue and pink dots denote type I (false hits) and type II (missed alarms) errors.

From the four plots it is clear that (a) is limited to a small learning database compared to the other three parametric methods. Despite their larger size, the training databases produced via MGC an MGD do not follow well the characteristics of the original historical data, thus excluding from the DT training process, areas of the state-space that do arise during testing. This leads to classification errors and a deterioration of predictive performance. The C-Vine-trained

DT covers well the test domain while also extrapolating the dependence structure and leading to exploration of the state-space at high density. A characteristic example of this is the high density of error in the yellow circle area for methods (a), (b) and (c) due to lack of training. On the other hand, the DT trained on C-Vine data has been trained in this area (hence the 'whiter' background) and carries out successful predictions. It is clear that an important advantage of the proposed model is that it can be sampled at arbitrarily high densities. For example, $\mathcal{Z}_k$ contains 0.9N observations, while a parametric model fitted to $\mathcal{Z}_k$ generates in our case 40,000 samples resulting in a higher concentration of points in the vicinity of the target decision boundary. In general, a larger training set can result in better classifier performance if over-fitting is avoided [38].

VII. Conclusion

In this research we show the potential for using data-driven proxies to complex problems and highlight the impact of the training database on proxy quality. To this end, we propose a general-purpose high-dimensional data modelling workflow comprising of clustering, dimension reduction and vine copulas. Using visual and statistical tests, the proposed copula model is shown to capture highly-complex dependence structures more accurately than conventional approaches. Through a case study on the 118-bus system we demonstrate that high-density sampling of the proposed model can result in superior proxies to describe the system's security boundary by improving diversity of the training set and moving beyond the limits of methods that rely exclusively on past observations.

The presented research contributes in enabling the shift from the current deterministic analysis paradigm to Monte Carlo frameworks. Future work will focus on developing intelligent sampling strategies that build on the identified model to improve learning rates of surrogate models. In addition, we aim to examine the potential for using model-free approaches such as Generative Adversarial Networks.

References

[1] P. Panciatici, G. Bareux, and L. Wehenkel, "Operating in the fog: security management under uncertainty," *IEEE Power Energy Mag.*, vol. 10, no. 5, pp. 40–49, Sep./Oct. 2012.

[2] N. Hargreaves, G. Taylor, A. Carter, and A. McMorran, "Developing emerging standards for power system data exchange to enable interoperable and scalable operational modeling and analysis," in *Proc. 46th Int. Universities′ Power Eng. Conf.*, 2011, pp. 1–5.

[3] L. Wehenkel, M. Pavella, E. Euxibie, and B. Heilbronn, "Decision tree based transient stability method a case study," *IEEE Trans. Power Syst.*, vol. 9, no. 1, pp. 459–469, Feb. 1994.

[4] L. Wehenkel, T. van Cutsem, and M. Pavella, "An artificial intelligence framework for on-line transient stability assessment of power systems," *IEEE Trans. Power Syst.*, vol. 4, no. 2, May 1989.

[5] C. Liu *et al.*, "A systematic approach for dynamic security assessment and the corresponding preventive control scheme based on decision trees," *IEEE Trans. Power Syst.*, vol. 29, no. 2, Mar. 2014.

[6] J. L. Cremer, I. Konstantelos, S. H. Tindemans, and G. Strbac, "Sample-derived disjunctive rules for secure power system operations," in *Proc. 2018 IEEE Int. Conf. Probabilistic Methods Applied to Power Systems (PMAPS)*, Jun. 24–28, 2018, Boise, ID, USA, doi: 10.1109/PMAPS.2018.8440373.

[7] T. Guo, S. Member, and J. V. Milanović, "Online identification of power system dynamic signature using PMU measurements and data mining," vol. 31, no. 3, pp. 1–9, 2015.

[8] L. Duchesne, "Machine learning of proxies for power systems reliability assessment," *M.S. thesis*, Dept. Elect. Eng. Comp. Sci., Univ. Liège, Liège, Belgium, 2016.

[9] V. Krishnan, J. D. McCalley, S. Henry, and S. Issad, "Efficient database generation for decision tree based power system security assessment," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 2319–2327, Nov. 2011.

[10] V. Krishnan and J. D. McCalley, "Importance sampling based intelligent test set generation for validating operating rules used in power system operational planning," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 2222–2231, Aug. 2013.

[11] E. C. Brechmann, C. Czado, and K. Aas, "Truncated regular vines in high dimensions with application to financial data," *Can. J. Statist.*, vol. 40, no. 1, pp. 68–85, 2012.

[12] C. Schoelzel and P. Friederichs, "Multivariate non-normally distributed random variables in climate research–introduction to the copula approach," *Nonlinear Processes Geophys.*, vol. 15, pp. 761–772, 2008.

[13] M. Sun, I. Konstantelos, and G. Strbac, "C-Vine copula mixture model for clustering of residential electrical load pattern data," *IEEE Trans. Power Syst.*, vol. 32, no. 3, pp. 2382–2393, May 2017.

[14] G. Papaefthymiou and D. Kurowicka, "Using copulas for modeling stochastic dependence in power system uncertainty analysis," *IEEE Trans. Power Syst.*, vol. 24, no. 1, pp. 40–49, Feb. 2009.

[15] M. Sun, I. Konstantelos, and G. Strbac, "Transmission network expansion planning with stochastic multivariate load and wind modelling," in *Proc. Int. Conf. Probablistic Methods Appl. Power Syst.*, Beijing, 2016, pp. 1–7.

[16] W. Hu, Y. Min, Y. Zhou, and Q. Lu, "Wind power forecasting errors modelling approach considering temporal and spatial dependence," *J. Modern Power Syst. Clean Energy*, vol. 5, no. 3, pp. 489–498, 2017.

[17] O. Grothe and J. Schnieders, "Spatial dependence in wind and optimal wind power allocation: A copula-based analysis," *Energy Policy*, vol. 39, no. 9, pp. 4742–4752, Sep. 2011.

[18] A. Lojowska, D. Kurowicka, G. Papaefthymiou, and L. Van der Sluis, "Stochastic modeling of power demand due to EVs using copula," *IEEE Trans. Power Syst.*, vol. 27, no. 4, pp. 1960–1968, Nov. 2012.

[19] D. Michel, D. Dupuis, and S. Couture-Cardinal, "Complementarity of hydro and wind power: Improving the risk profile of energy inflows," *Energy Policy*, vol. 37, no. 12, pp. 5376–5384, 2009.

[20] S. Sriboonchitta, J. Liu, V. Kreinovich, and H. T. Nguyen, "Vine copulas as a way to describe and analyze multi-variate dependence in econometrics: computational motivation and comparison with Bayesian networks and fuzzy approaches," in *Modeling Dependence in Econometrics*. Cham, Switzerland: Springer, 2014, pp. 169–184.

[21] I. Goodfellow *et al.*, "Generative adversarial nets," *Adv. Neural Inf. Proc. Syst.*, pp. 2672–2680, 2014.

[22] Y. Chen, Y. Wang, D. Kirschen, and B. Zhang, "Model-free renewable scenario generation using generative adversarial networks," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3265–3275, May 2018.

[23] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for GANs do actually Converge?" *Proc. Mach. Learn. Res.*, vol. 80, pp. 3481–3490, 2018.

[24] S. Arora and Y. Zhang, "Do GANs actually learn the distribution? An empirical study," unpublished paper, 2017. [Online]. Available: https://arxiv.org/abs/1706.08224v1

[25] M. Sun, I. Konstantelos, and G. Strbac, "Evaluating composite approaches to modelling high-dimensional stochastic variables in power systems," in *Proc. Power Syst. Comput. Conf.*, Genoa, 2016.

[26] M. H. Vasconcelos *et al.*, “Online security assessment with load and renewable generation uncertainty: The iTesla project approach,” in *Proc. Int. Conf. Probabilistic Methods Appl. Power Syst.*, Beijing, China, 2016, pp. 1–8.

[27] I. Konstantelos *et al.*, “Implementation of a massively parallel dynamic security assessment platform for large-scale grids,” *IEEE Trans. Smart Grid*, vol. 8, no. 3, pp. 1417–1426, May 2017.

[28] R. Tibshirani, W. Guenther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” *J. Roy. Statistical Soc., B*, vol. 63, no. 2, pp. 411–423, 2001.

[29] J. L. Myers and A. D. Well, *Research Design and Statistical Analysis*, 2nd ed. New York, NY, USA: Lawrence Erlbaum, 2003, p. 508.

[30] H. Joe, “Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters,” *Distribution Fixed Marginals Related Topics*, vol. 28, pp. 120–141, 1996.

[31] J. Mai and M. Scherer, *Simulating Copulas: Stochastic Models, Sampling Algorithms and Applications*. Singapore: World Scientific, 2012.

[32] T. Bedford and R. M. Cooke, “Probability density decomposition for conditionally dependent random variables modeled by vines,” *Ann. Math. Artif. Intell.*, vol. 32, pp. 245–268, 2001.

[33] K. Aas, C. Czado, A. Frigessi, and H. Bakken, “Pair-copula constructions of multiple dependence,” *Insurance, Math. Economics*, vol. 44, no. 2, pp. 182–198, 2009.

[34] D. Kurowicka and R. M. Cooke, *Uncertainty Analysis with High Dimension Dependence Modeling*. Chichester, U.K.: Wiley, 2012.

[35] T. Bedford and R. M. Cooke, “Vines–A new graphical model for dependent random variables,” *Ann. Statist.*, vol. 30, no. 4, pp. 1031–1068, 2002.

[36] Illinois Institute of Technology (IIT), IEEE 118-bus System Data, Elect. Comput. Eng. Dept., 2015. [Online]. Available: http://motor.ece.iit.edu/Data/

[37] P. Pinson and R. Girard, “Evaluating the quality of scenarios of short-term wind power generation,” *Appl. Energy*, vol. 96, pp. 12–20, 2012.

[38] J. Morgan, R. H. Daugherty, R. Hilchie, and B. Carey, “Sample size and modeling accuracy of decision tree based data mining tools,” *Acad. Inf. Manage. Sci. J.*, vol. 6, no. 2, pp. 71–99, 2003.

[39] J. A. Doornik and H. Hansen, “An omnibus test for univariate and multivariate normality,” *Oxford Bull. Economics Statist.*, vol. 70, pp. 927–939, 2008.

[40] Github.com, Open source truncated C-Vine toolbox, 2017. [Online]. Available: https://github.com/itesla/ipst/tree/master/sampling, Accessed on: Aug. 30, 2017.

[41] A. Sklar, “Fonctions de répartition à n dimensions et leurs marges,” *Publication Inst. Statist. Univ. Paris*, vol. 8, 229–231, 1959.

[42] S. S. Shapiro and M. B. Wilk, “An analysis of variance test for normality (complete samples),” *Biometrika*, vol. 52, no. 4, pp. 591–611, 1965.

[43] F. J. Massey, “The Kolmogorov–Smirnov test for goodness of fit,” *J. Amer. Statist. Assoc.*, vol. 46, no. 253, pp. 68–78, Mar. 1951.

[44] B. Aslan and G. Zech, “New test for the multivariate twosample problem based on the concept of minimum energy,” *J. Statistical Comput. Simul.*, vol. 75, no. 2, pp. 109–119, 2005.

Ioannis Konstantelos (M’12) received the M.Eng. degree in electrical and electronic engineering from Imperial College London, London, U.K., in 2007, and the Ph.D. degree from the same university in 2013 in the field of electrical energy systems. His research interests include mathematical programming and machine learning techniques applied to the planning and operation of energy systems.

Mingyang Sun (M’16) received the Ph.D. degree from Imperial College London, London, U.K., in 2017. He is currently a Research Associate with Imperial College London. His research focuses on big data analytics in power systems.

Simon H. Tindemans (M’13) received the M.Sc. degree in physics from the University of Amsterdam, Amsterdam, The Netherlands, in 2004, and the Ph.D. degree from Wageningen University, Wageningen, The Netherlands, in 2009. From 2010–2017, he was a Postdoctoral Researcher with the Control and Power Research Group, Imperial College London, London, U.K. Since 2018, he is an Assistant Professor with the Delft University of Technology, Delft, The Netherlands. His research interests include computational methods for power system reliability assessment, statistical learning, and stochastic control for demand response.

Samir Issad (M’09) received the graduate degree from the French High Engineering School, Ecole des Mines de Nancy, Nancy, France, in 2007 and the master’s degree in mathematics from the French Université Henri-Poincaré, Nancy, France, in 2007. He joined Research and Development Department, Reseau De Transport D Electricite, Versailles, France, in 2007, working on power system reliability studies using statistical and probabilistic methods.

Patrick Panciatici (SM’11) received the Ingenieur degree from Supelec, Gif-sur-Yvette, France, in 1984. He joined Electricite de France S.A., France, Reseau de transport d’ electricite (RTE), France, Research and Development (R&D), in 1985, then RTE (French Transmission System Operator) in 2003 and participated in the creation of the Internal R&D Department. He has more than 30 years’ experience in the field of R&D for transmission systems. He is currently a Scientific Advisor, and coordinates and supervises long-term research activities with the R&D Department, RTE. He is the Vice-Chair of the Power Systems Engineering Research Center Industrial Advisory Board. He is a member of the CIGRE.

Goran Strbac (M’95) is a Professor of electrical energy systems with Imperial College, London, London, U.K. His current research interests include electricity generation, transmission and distribution operation, planning and pricing, and integration of renewable and distributed generation in electricity systems.