



A copula-based data augmentation strategy for the sensitivity analysis of extreme operational losses

A. Khorrami Chokami & G. Rabitti

To cite this article: A. Khorrami Chokami & G. Rabitti (2025) A copula-based data augmentation strategy for the sensitivity analysis of extreme operational losses, Quantitative Finance, 25:5, 841-849, DOI: [10.1080/14697688.2025.2487103](https://doi.org/10.1080/14697688.2025.2487103)

To link to this article: <https://doi.org/10.1080/14697688.2025.2487103>



© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 23 Apr 2025.



[Submit your article to this journal](#)



Article views: 1314



[View related articles](#)



[View Crossmark data](#)



Citing articles: 1 [View citing articles](#)

A copula-based data augmentation strategy for the sensitivity analysis of extreme operational losses

A. KHORRAMI CHOKAMI[†] and G. RABITTI^{*‡}

[†]Department of Mathematics and Computer Science, Università di Cagliari, Cagliari, Italy

[‡]Department of Actuarial Mathematics and Statistics, Heriot-Watt University and Maxwell Institute for Mathematical Sciences, Edinburgh, UK

(Received 6 August 2024; accepted 26 March 2025)

In this work, we aim to assess the importance of macroeconomic and financial variables for operational losses of UniCredit Bank. To achieve this, we consider the Shapley effects as a variance-based measure of importance. However, the small number of observations of extreme losses makes the estimation of the Shapley effects challenging. To address this issue, we proposed augmenting the sample of extreme observations using vine copulas and calculating the Shapley effects on the augmented sample. The effectiveness of this procedure is supported by a numerical simulation. Findings obtained with our methodology applied to the UniCredit Bank data show its usefulness for the risk management of operational losses.

Keywords: Operational risk; Extreme values; Global sensitivity analysis; Vine copulas

1. Introduction

Operational risk denotes the risk of losses due to failed internal processes, people, systems, or external events. This type of risk can significantly impact insurance companies, banks and financial institutions. Consequently, it is important to understand the main drivers of operational risk losses for proper and effective risk management (Chernobai *et al.* 2011, Cope *et al.* 2012, Chavez-Demoulin *et al.* 2016). This insight can be obtained using sensitivity analysis, which denotes a set of techniques to determine which variables most significantly influence the model output, i.e. the operational losses. For example, in the financial literature, Bolker *et al.* (2021) use sensitivity analysis on climate and economic models, and more recently, Borgonovo *et al.* (2024) apply it to insurance internal models.

There are several methods for sensitivity analysis of the model output (Borgonovo and Plischke 2016). A traditional approach is based on the variance decomposition (Sobol 1993, Saltelli *et al.* 2008), which aims to attribute the explained variance to each subset of variables. Another approach involves moment-independent methods, which assess how the distribution of the quantity of interest changes when conditioned on specific variables. Within this framework are the Borgonovo measure (Borgonovo 2007), the

Cramer–Von Mises sensitivity measure (Gamboa *et al.* 2018), and the recent sensitivity measures based on optimal transport (Borgonovo *et al.* 2024). Among these sensitivity methods, a measure of importance that is attracting increasing interest are the Shapley effects, which allocate the variance of the quantity of interest according to the contribution of each individual variable (Owen 2014, Song *et al.* 2016, Owen and Prieur 2017, Iooss and Prieur 2019). Compared to other variance-based sensitivity methods, the sum of the Shapley effects equals the total variance, regardless of the joint distribution of the variables. This property makes the interpretation of these measures particularly intuitive. In the context of operational losses, an explanatory variable is considered important if it accounts for a high fraction of the operational loss variance. This approach allows for the ranking of the covariates, which is useful for risk management decisions. Specifically, if the focus is on extreme losses, Il Idrissi *et al.* (2021) propose using Shapley effects to allocate the tail variance.

An open issue concerns the estimation of the Shapley effects. Recently, Mase *et al.* (2019) have proposed a procedure based on cohorts of similar observations that does not require the fitting of any predictive model (see also Vallarino *et al.* 2023 for an application of this technique to insurance rate-making). However, in the case of extreme operational losses, there is typically a limited number of observations

*Corresponding author. Email: G.Rabitti@hw.ac.uk

available. Consequently, this procedure might become inaccurate because the variance explained by subsets of variables is estimated with few data points.

To deal with this issue, we propose in this article a data augmentation strategy based on multivariate copulas used to capture the dependence structure in the data. Vine copulas provide a highly flexible approach for modeling general multivariate dependency structures, making them a more effective tool compared to other classes of copulas, particularly when the number of variables is high (Bedford and Cooke 2002, Czado 2019, Czado and Nagler 2022). They are based on specific factorizations of the joint density in terms of bivariate copulas aggregated in a tree-like structure. Due to their versatility, vine copulas have been used in the financial literature to capture multivariate dependence for multivariate statistical arbitrage strategies (Stübinger *et al.* 2018) and for risk management purposes (Brechmann and Czado 2013).

In this work, we use for the first time vine copulas with the purpose of improving the accuracy of the estimates of the Shapley effects. Precisely, by simulating new observations from the selected vine copula applied to the extreme operational loss dataset, we can increase the sample size while maintaining the original dependence structure between the operational losses and the covariates. The non-parametric approach that we propose is very flexible due to the versatility of vine copulas, and therefore it avoids the need to specify predictive models for the tails. We remark that this data augmentation strategy can also be useful in the estimation of other importance measures, such as moment-independent ones. In our numerical simulation, we illustrate that the estimates of the Shapley effects after the data augmentation are more reliable.

This paper is structured as follows. Section 2 presents the Shapley effects. Section 3 contains an overview of the vine copula models. In Section 4, we illustrate the procedure with a numerical simulation. In Section 5, we apply the non-parametric method to the UniCredit Bank operational loss data considered in Hambuckers *et al.* (2018) and Bee *et al.* (2019, 2021).

2. Variable importance using Shapley effects

The Shapley effects constitute an important sensitivity measure based on the Shapley value (Shapley 1953). The Shapley value stems from the cooperative game theory literature as a solution to the problem of allocating the total payoff of a game to each player in a coalition, according to their contribution.

Formally, consider a set of players $D = \{1, 2, \dots, d\}$ and let $u \subseteq D$ be any possible subcoalition. Let $\text{val} : 2^D \rightarrow \mathbb{R}$ be the value function of the game. It must satisfy the condition $\text{val}(\emptyset) = 0$, meaning that no value is generated when there are no players. The Shapley value of the j th player is expressed as

$$\phi_j = \sum_{u \subseteq D \setminus \{j\}} \frac{(d - |u| - 1)!|u|!}{d!} (\text{val}(u \cup \{j\}) - \text{val}(u)), \quad (1)$$

where $|u|$ denotes the cardinality of the set u . We observe that the Shapley value is based on the marginal increase

$\text{val}(u \cup \{j\}) - \text{val}(u)$, measuring the change of the value produced by the coalition u when the player j joins, averaged over all possible coalitions. Shapley (1953) proved that the allocation method in equation (1) is the only one satisfying the following properties:

Efficiency: The sum of Shapley values of all players equates to the total value of the game $\sum_{j=1}^d \phi_j = \text{val}(D)$.

Symmetry: If players i and j make the same marginal contribution to any coalition, $\text{val}(u \cup \{i\}) = \text{val}(u \cup \{j\})$ for all $u \subseteq D$ with $i, j \notin D$, then $\phi_i = \phi_j$.

Dummy: If j is a dummy player, $\text{val}(u \cup \{j\}) = \text{val}(u)$ for all $u \subseteq D$, then $\phi_j = 0$.

Linearity: If val and val' have Shapley values ϕ and ϕ' respectively, then the game with value $\text{val}(u) + \text{val}'(u)$ has Shapley value $\phi_j + \phi'_j$ for all $j \in D$.

Various value functions have been considered in the literature (Borgonovo *et al.* 2024). In the context of variable importance, Owen (2014) suggests to identify covariates with players and their explained variance as the value function. Consider the model $Y = g(X_1, X_2, \dots, X_d)$, where $g \in L^2$ and the variables $X_1, X_2, \dots, X_d$ are independent. Under these assumptions, the total variance of Y can be decomposed according to the Functional Analysis of Variance (ANOVA) as

$$\text{Var}(Y) = \sum_{j=1}^d V_j + \sum_{i < j} V_{ij} + \dots + V_{12\dots d}, \quad (2)$$

where the variance components are recursively defined as $V_j = \text{Var}(\mathbb{E}[Y | X_j])$, $V_{ij} = \text{Var}(\mathbb{E}[Y | X_i, X_j]) - V_i - V_j$ and so on. Each variance term V_j is called the main effect of the variable X_j and represents the variance explained by this variable, for any $j = 1, \dots, d$. The terms V_u represent the variance explained by the interaction among the variables in the group u . The (closed) Sobol' effects (Sobol' 1993) of the group u are defined as $\text{Var}(\mathbb{E}[Y | X_u])$ and they are equal to $\sum_{v \subseteq u} V_v$. Such values quantify the contribution of the variables in the group u to the overall variance of Y . However, the variance decomposition in equation (2) does not hold anymore when variables are dependent due to covariance terms (Owen and Prieur 2017). Nonetheless, Owen (2014) considers as value function the Sobol' effects, i.e.

$$\text{val}(u) = \text{Var}(\mathbb{E}[Y | X_u]). \quad (3)$$

The Shapley values computed with this value function are called Shapley effects (Song *et al.* 2016), and enjoy these appealing properties:

- (i) $\phi_j \geq 0$ for all $j = 1, 2, \dots, d$. This holds because $\text{val}(u \cup \{j\}) - \text{val}(u)$ is positive as the variance explained increases when a new covariate is considered for all $u \subseteq D$.
- (ii) The sum $\sum_{j=1}^d \phi_j$ equals to the variance of Y . This follows from the efficiency property, meaning that we allocate the variance of the model output.
- (iii) They remain well-defined under any type of dependency among the variables (unlike Sobol' effects).

Hence, the covariate with highest Shapley effect explains the highest fraction of the variance and is deemed as the most important. Conversely, a covariate with null Shapley effect can be considered as non-influential. From a mathematical perspective, using Shapley effects bypasses the problem of having to generalize equation (2) to the case of dependent variables, as observed in Owen and Prieur (2017).

In general, the estimation of the Shapley values is computationally burdensome for datasets with even moderate number of covariates, since the number of value functions to be estimated in equation (1) is 2^d . To compute the 2^d value functions defined in equation (3) to obtain the Shapley effects, we consider the approach proposed in Mase *et al.* (2019). This method is based on sets of observations sharing similar values for any sub-group u , called cohorts. Mase *et al.* (2019) show that the sample means of these cohorts are an estimate of the conditional means appearing in equation (3). This empirical estimation strategy allows them to estimate the variance contributions directly from a dataset without the need to estimate g . However, the estimates of the conditional variances become less accurate when considering only extreme observations. Hence, in the next section, we propose a data augmentation strategy with which we can simulate new observations that have a similar dependency structure to that of the observed extreme loss data.

3. Data augmentation with vine copulas

Copulas are fundamental tools for modeling the dependence structure between variables in statistical models. More specifically, a copula is a multivariate distribution function with standard uniform margins. Thanks to the Theorem of Sklar, any multivariate joint distribution function can be decomposed into its marginal distribution functions and a copula, which captures the dependence structure regardless of the marginals. Precisely, let us denote with C the $(d + 1)$ -dimensional copula associated to the multivariate continuous distribution $F(y, x_1, x_2, \dots, x_d)$. Then the density corresponding to F is

$$\begin{aligned} f(y, x_1, \dots, x_d) \\ = c(F_Y(y), F_1(x_1), \dots, F_d(x_d)) f_Y(y) f_1(x_1) \cdots f_d(x_d), \end{aligned}$$

where c is the copula density. Hence, a flexible method is needed to express c , capturing the dependence among the operational losses and the explanatory variables.

Vine copulas have emerged as a popular choice due to their ability to capture complex dependence patterns that may not be adequately captured by other copula models (Czado and Nagler 2022). These models consist into breaking down the joint distribution into simpler conditional relationships, which are more manageable and interpretable. Vine copulas consist of a sequence of bivariate copulas arranged in a tree-like structure. Formally, adopting the notation in Czado and Nagler (2022), we recall that a tree is a connected acyclic graph $T = (N, E)$ with node set N and edge set E . We want to factorize the joint distribution of the vector $(Y, X_1, \dots, X_d)$ of dimension $d + 1$, corresponding to the operational loss Y together with the risk factors $X_1, \dots, X_d$. To simplify

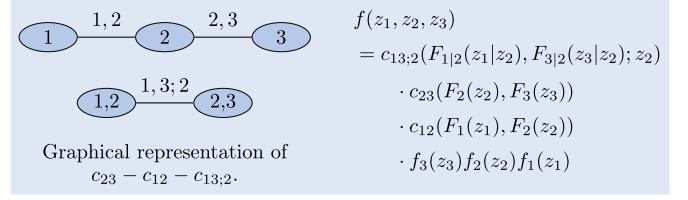


Figure 1. A possible copula representation for a 3-variate pdf (Czado and Nagler 2022). The term $c_{13;2}$ denotes the conditional copula density between variables Z_1 and Z_3 given Z_2 .

the notation of the tree structure, denote $(Y, X_1, \dots, X_d) := (Z_1, Z_2, \dots, Z_{d+1})$. We provide an illustration of a factorization of a 3-variate density in figure 1. Figure 1 displays two trees, which indicate how to split the joint density into conditional and marginal terms. Intuitively, the latter terms corresponding to the nodes of the upper tree, while the former terms representing the pair copula densities correspond to the nodes and the edges of the lower graph. The intuition provided in figure 1 is formalized below for the general case, with the use of a *regular vine tree structure* (Bedford and Cooke 2002). This structure comprises interconnected trees, where the edges of one tree transition into the nodes of the subsequent one. This graphical representation aids in visualizing and understanding the complex interdependencies among multiple variables.

When the dimension of the copula is greater than 5, the subgraphs are no longer constrained to paths or stars, leading to a significant increase in the number of potential decompositions.

The set of graphs $\mathcal{V} = (T_1, \dots, T_d)$ is a regular vine tree sequence on $d + 1$ elements if:

- (i) T_1 is a tree with node set $N_1 = \{1, \dots, d + 1\}$ and edge set E_1 .
- (ii) For $k \geq 2$, T_k is a tree with node set $N_k = E_{k-1}$ and edge set E_k .
- (iii) For $k = 2, \dots, d$ and $\{a, b\} \in E_k$ it must hold that $|a \cap b| = 1$. It states that an edge between nodes in tree T_k is only possible, if the corresponding edges in T_{k-1} share a common node.

The following symbols are needed to represent the relation between the regular tree structure and the factorization of the target multivariate density function. For any edge $e \in E_\ell$ define the set

$$\begin{aligned} A_e &:= \{k \in N_1 \mid \exists e_1 \in E_1, \dots, e_{\ell-1} \in E_{\ell-1} \text{ such that} \\ &\quad k \in e_1 \in \dots \in e_{\ell-1} \in e\} \end{aligned}$$

The set A_e is called the complete union of the edge e . The conditioning set D_e of an edge $e = \{a, b\}$ is defined by $D_e := A_a \cap A_b$ and the conditioned sets $C_{e,a}$ and $C_{e,b}$ are given by

$$C_{e,a} := A_a \setminus D_e, \quad C_{e,b} := A_b \setminus D_e \quad \text{and} \quad C_e := C_{e,a} \cup C_{e,b}.$$

We often abbreviate each edge $e = (C_{e,a}, C_{e,b}; D_e)$ in the vine tree sequence by $e = (e_a, e_b; D_e)$. The construction of the regular vine copula follows by specifying the triplet $(\mathcal{F}, \mathcal{V}, \mathcal{B})$ with:

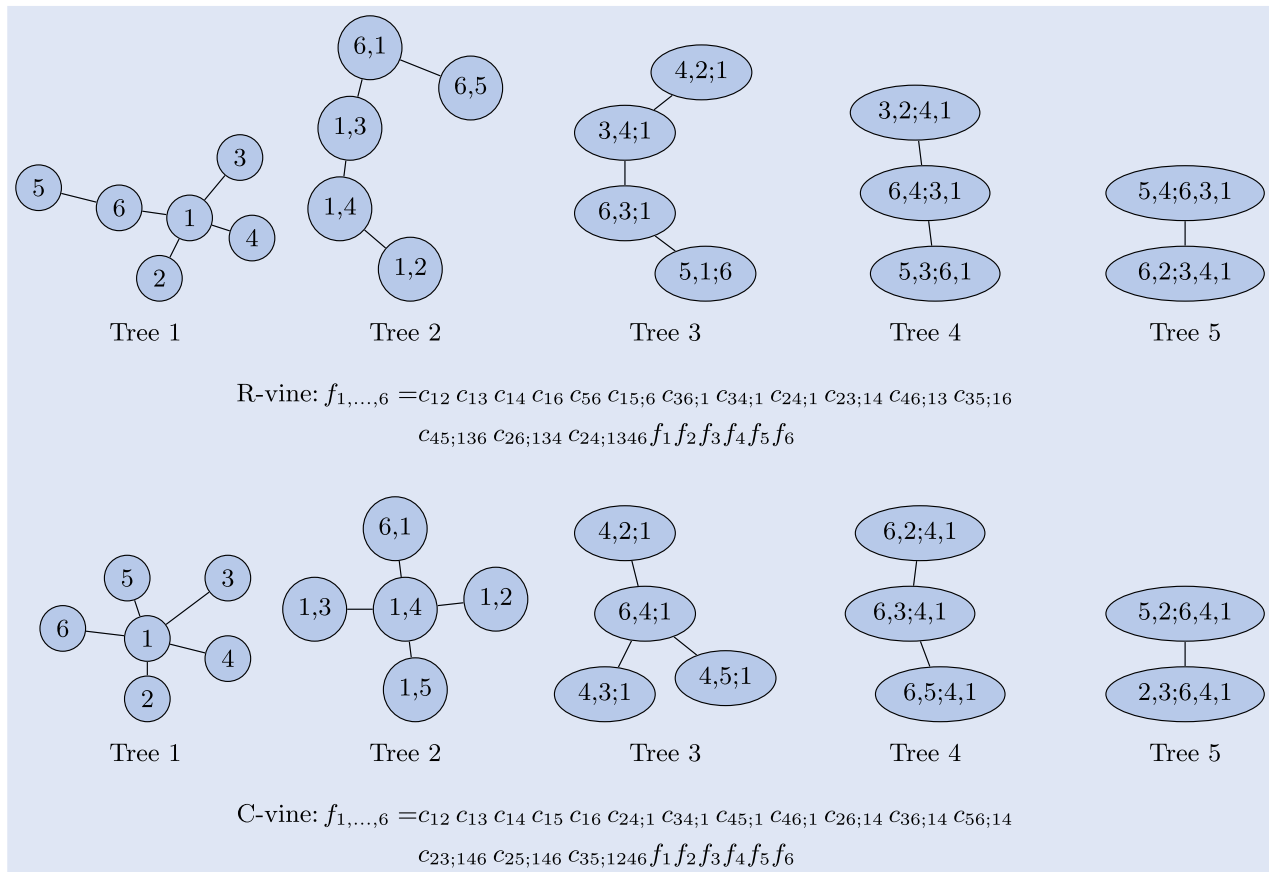


Figure 2. Tree sequences corresponding to an R-vine copula (upper part) and a C-vine copula (lower part), relative to two different factorizations (given below each graph) of a six-dimensional pdf (Czado and Nagler 2022). For readability, pdfs and pair copulas are written in a compact way. The labels of the trees indicate their position in the sequence $\mathcal{V}$.

- (1) Marginal distributions: $\mathcal{F} = (F_1, \dots, F_{d+1})$ is a vector of continuous marginal distribution functions of the random variables $Z_1, Z_2, \dots, Z_{d+1}$.
- (2) Regular vine tree sequence: $\mathcal{V}$ is an R-vine tree sequence on $d + 1$ elements.
- (3) Bivariate copulas: The set $\mathcal{B} = \{C_e \mid e \in E_\ell; \ell = 1, \dots, d\}$, where C_e is a bivariate copula with density. Here E_ℓ is the edge set of tree T_ℓ in the R-vine tree sequence $\mathcal{V}$.
- (4) Relation between R-vine tree sequence and the set of bivariate copulas: for each $e \in E_\ell, \ell = 1, \dots, d, e = \{a, b\}$, $C_e(\cdot, \cdot)$ is the copula associated with the conditional distribution of $Z_{e,a}$ and $Z_{e,b}$ given $\mathbf{Z}_{D_e}$.

In the literature, a regular vine copula is referred to as an R-vine (Czado and Nagler 2022). When all the trees are stars, the vine copula is known as a canonical vine or C-vine. An illustration of the difference between an R-vine and a C-vine in terms of tree structure is provided in figure 2, for a six-dimensional case. This figure shows the connection between a sequence of trees and the corresponding factorization of the multivariate pdf.

For the selection of the copula, it is convenient to work with families of parametric copulas. Considering again the three-dimensional example in figure 1, calling the parameter

vector $\boldsymbol{\theta} := (\boldsymbol{\theta}_{12}, \boldsymbol{\theta}_{23}, \boldsymbol{\theta}_{13;2})$, the factorization of the three-dimensional pdf is

$$\begin{aligned} f(z_1, z_2, z_3; \boldsymbol{\theta}) &= c_{13;2}(F_{1|2}(z_1 | z_2), F_{3|2}(z_3 | z_2); z_2, \boldsymbol{\theta}_{13;2}) \\ &\quad \cdot c_{23}(F_2(z_2), F_3(z_3); \boldsymbol{\theta}_{23}) \cdot c_{12}(F_1(z_1), F_2(z_2); \\ &\quad \boldsymbol{\theta}_{12}) \cdot f_3(z_3) f_2(z_2) f_1(z_1). \end{aligned}$$

Parameter estimation involves fitting vine copula models to observed data using maximum likelihood estimation. Model selection aims at choosing the appropriate vine structure and copula families based on the AIC (see Czado 2019 for more details). To augment the extreme loss sample with vine copulas, we perform the following steps:

- Step 1** Isolate the dependence structure by transforming the marginal distributions into the Uniform(0, 1) distribution through the rank transformation (McNeil et al. 2015)
- Step 2** Estimate the vine copula tree structure by maximum likelihood and select it based on the AIC with the R function `RVineStructureSelect` in the package `VineCopula`
- Step 3** Simulate new observation from the selected vine copula
- Step 4** Apply to the new data the empirical inverse cumulative distribution functions to recover the original scales of the marginals.

Table 1. Mean Squared Errors of the estimated Shapley effects for the linear model, considering different correlation coefficients among the variables, both with and without data augmentation (DA).

n	$\rho = 0$		$\rho = 0.5$		$\rho = 0.8$	
	No DA	With DA	No DA	With DA	No DA	With DA
50	0.0128	0.00318	0.0104	0.00301	0.00767	0.00363
250	0.00370	0.00163	0.00350	0.00210	0.00304	0.00208
500	0.00275	0.00158	0.00259	0.00168	0.00211	0.00184

Note: Each row corresponds to a different original sample size. With data augmentation, we simulated 2000 new values, and the procedure was repeated 20 times.

Table 2. Abbreviations, descriptions and tail sample sizes of event types.

Event type	Description	Tail observations
IFRAUD	Internal frauds	64
EFRAUD	External frauds, related to payments and others	320
EPWS	Employment practices and workplace safety	115
CPBP	Clients, products and business practices, related to derivatives, financial instruments and others	807
EDPM	Execution, delivery and process management, related to financial instruments, payments and others	661

Lastly, we note that there is a parametric alternative to Steps 1 and 4, if the analyst fits parametric distributions to the marginals. However, we do not adopt this procedure because it is less flexible and might introduce model error for the fitted marginals.

4. Numerical simulation

In this section, we apply the methodology previously described to an analytical test case (see Iooss and Prieur 2019). We sample $n = 50, 250, 500$ observations from the model $Y = 2X_1 + X_2 + 0.3X_3$, where $X_1 \sim \mathcal{N}(0, 4)$, $X_2 \sim \mathcal{N}(0, 1)$, $X_3 \sim \mathcal{N}(0, 2.25)$. The covariance matrix describing the dependence structure is

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & \rho\sigma_2\sigma_3 \\ 0 & \rho\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix},$$

where ρ takes values 0, 0.5 and 0.8. Then we apply Steps 1–4 to simulate 2000 observations for the computation of the Shapley effects, where the variables X_1 , X_2 and X_3 are seen as the players in equation (1), with value function in equation (3) divided by the variance. This normalization allows us to interpret the Shapley effects as fractions of the explained variance (note that the results of the analysis remain invariant). Given

$$\sigma_{\text{tot}}^2 = \beta_1^2\sigma_1^2 + \beta_2^2\sigma_2^2 + \beta_3^2\sigma_3^2 + 2\beta_2\beta_3\sigma_2\sigma_3,$$

the true values of the normalized Shapley effects in this case are (Iooss and Prieur 2019)

$$\begin{aligned} \phi_1 &= \frac{\beta_1^2\sigma_1^2}{\sigma_{\text{tot}}^2}, \\ \phi_2 &= \frac{\beta_2^2\sigma_2^2 + \beta_2\beta_3\sigma_2\sigma_3 + \frac{1}{2}(\beta_3^2\sigma_3^2 - \beta_2^2\sigma_2^2)}{\sigma_{\text{tot}}^2}, \\ \phi_3 &= \frac{\beta_3^2\sigma_3^2 + \beta_2\beta_3\sigma_2\sigma_3 + \frac{1}{2}(\beta_2^2\sigma_2^2 - \beta_3^2\sigma_3^2)}{\sigma_{\text{tot}}^2}. \end{aligned}$$

We have repeated this procedure 20 times. The Mean Square Errors are displayed in table 1. From table 1, we note that both approaches produce Shapley effects that become closer to the true ones as the original sample increases, as the error decreases. However, with data augmentation, we note a reduction in error across all cases. Concerning the computational time, producing table 1 required approximately 9 min on a standard laptop with 8GB of RAM.

5. Application

We have considered the data from the UniCredit operational risk department, analyzed in Hambuckers *et al.* (2018) and Bee *et al.* (2019, 2021). Losses are classified into five risk categories, described in table 2.

In this work, we consider extreme operational losses as those that are above the 95-th percentile of the respective loss type distribution. The corresponding sample sizes are reported in the last column of table 2. There are 20 explanatory variables for these operational losses, represented in table 3.

We select the vine copula type with the lowest AIC. Results are reported in table 4. It is possible to see that the R-vine copulas are always selected, except for the CPBP event type. After having estimated the dependence structure, we proceed with Steps 3 and 4 to obtain 2000 new observations for each operational event type. We remark that we could have used a larger sample size n . However, the computational complexity of the cohort approach of Mase *et al.* (2019) is $\mathcal{O}(n^2 \cdot 2^d)$ (Rabitti and Tzougas 2025). This implies that, for a fixed dimension d , the sample size remarkably affects the computational time. In our application, with 2000 observations and 20 variables, the time required to estimate the Shapley effects for each loss type is approximately 2 days. Therefore, analysts should evaluate the available computational budget

Table 3. Summary of the explanatory variables.

Variable	Category	Variable	Category
Leverage ratio (LR)	Firm-specific	Monetary aggregate M1 growth rate (M1)	Macroeconomic
Tier-I capital ratio (TCR)	Firm-specific	Consumer loans rate < 1 year in Italy (LOR IT)	Macroeconomic
% revenue coming from fees (PRF)	Firm-specific	Consumer loans rate < 1 year in EU (LOR EU)	Macroeconomic
Deposit growth rate (DGR)	Firm-specific	S&P 500 returns	Financial
UniCredit stock returns (SR)	Firm-specific	TR EU Stock Index returns (TRSI)	Financial
Italian unemployment rate (UR IT)	Macroeconomic	FTSE MIB index returns (MIB)	Financial
EU unemployment rate (UR EU)	Macroeconomic	VIX	Financial
Italian GDP growth rate (GDP IT)	Macroeconomic	VFTSE	Financial
EU GDP growth rate (GDP EU)	Macroeconomic	3-month Italian interbank rate (STratelag)	Financial
EU housing price growth rate (HPI)	Macroeconomic	10-year Italian government bond yield (Ltratelag)	Financial

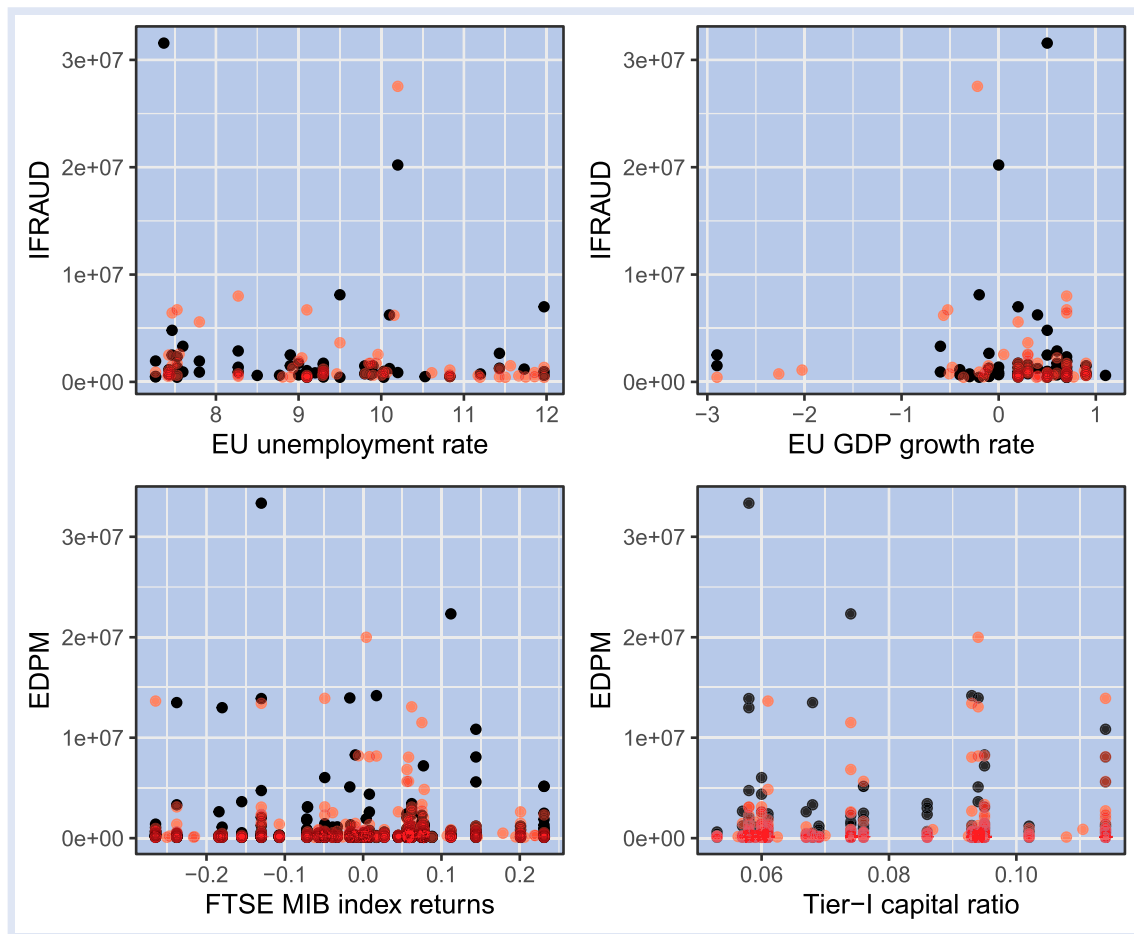


Figure 3. Scatterplots comparing real and simulated data for the dependence between two event types and some covariates. The black dots represent the real data, while the red dots represent the simulated data generated using the selected vine copula. In each panel, the number of simulated observations matches the number of original observations for visualization purposes.

and time constraints relative to the specific application under investigation.

To demonstrate the adequacy of the simulated data compared to the real data, we have plotted an illustrative example in Figure 3.

We have then applied the approach of Mase *et al.* (2019) to compute the Shapley values. The computational time was heavily influenced by the number of variables, due to the well-known exponential cost of the Shapley values. With the chosen sample dimension, the computational time was approximately 30h for each loss type. For this reason, we have not selected a larger sample. Figure 4 illustrates the obtained Shapley effects.

We can compare the findings provided by figure 4 with those of Hambuckers *et al.* (2018), who perform a Pareto regression on the same dataset and select variables according to their statistical significance (i.e. small p -values). Overall, our study identifies GDP, prflag, and the S&P 500 index as general drivers of financial losses, with significant individual variability for specific loss types. This is consistent with Abdymomunov *et al.* (2020), who provide evidence that macroeconomic variables are important drivers of operational risks. Contrary to the findings of Hambuckers *et al.* (2018), we discovered that the VIX financial index is generally not a significant driver of operational losses. Beyond the three general drivers (GDP, prflag, and S&P 500), there

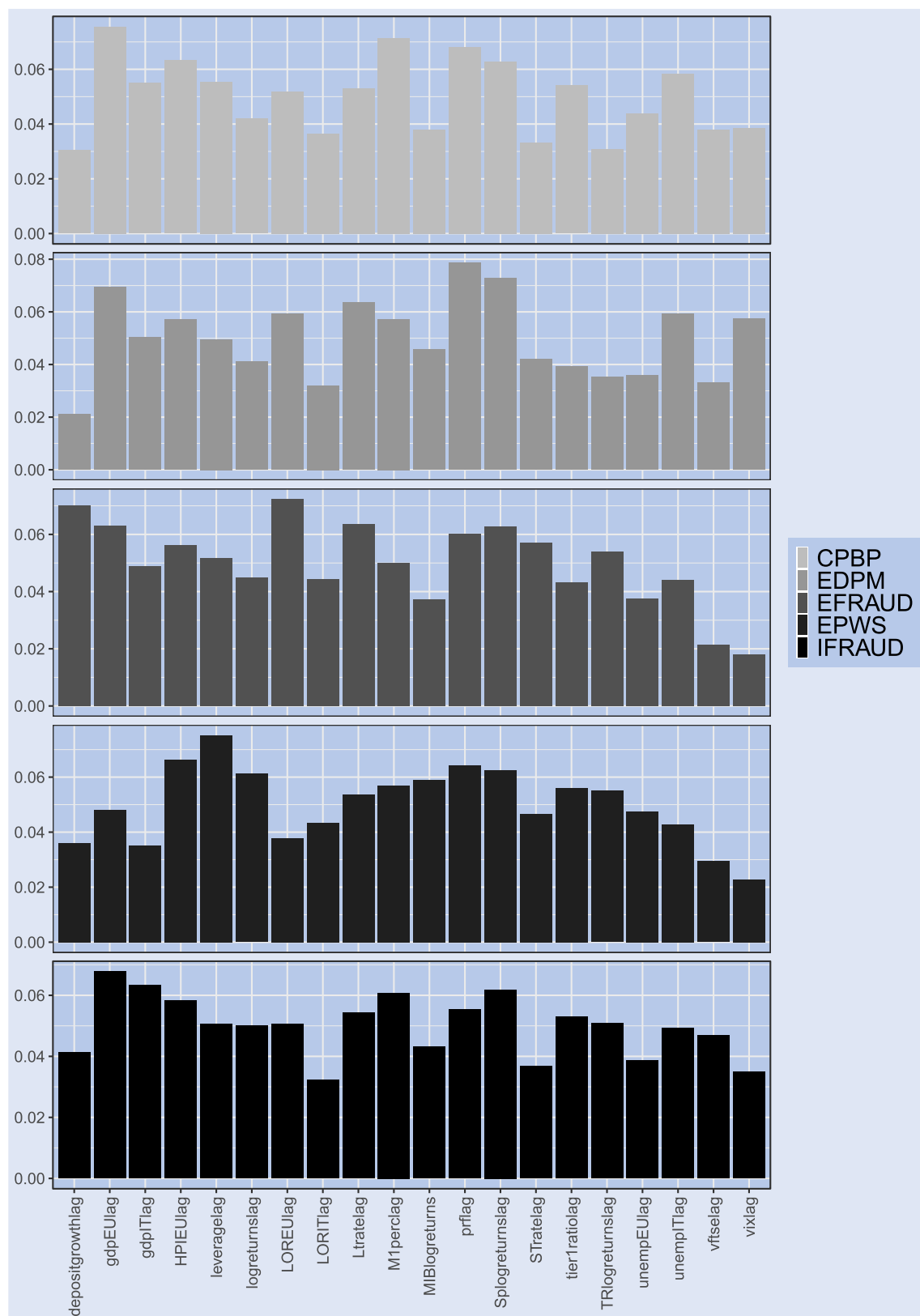


Figure 4. Barplots representing the normalized Shapley effects of each covariate for every operational loss.

is considerable individual variability in what influences specific types of losses. Concerning the CPBP losses, they are driven by macroeconomic (GDP, M1 growth rate), financial (S&P 500), and firm-specific factors (prflag), consistently with Hambuckers *et al.* (2018). This underscores the impact

of macroeconomic stability and firm performance on derivative and financial product losses. Regarding the EDPM and CPBP losses, Hambuckers *et al.* (2018) consider them to be similar, also influenced by the same drivers. However, we find that the loan rate (Ltrate) is particularly important for EDPM.

Table 4. Comparison of AIC values for different event types.

Event Type	AIC R-vine	AIC C-vine
IFRAUD	− 2552.84	− 2251.31
EFRAUD	− 13086.57	− 12687.28
EPWS	− 4686.59	− 4500.38
CPBP	− 34722.15	− 34812.06
EDPM	− 28964.51	− 27954.43

Unlike CPBP and EDPM, EFRAUD shows different patterns of importance. While GDP and deposit growth rate are important as in Hambuckers *et al.* (2018), the lagged loan rate (LOREUlag) emerges as the most critical variable in our analysis. Additionally, Ltrate is important. Taken together, these covariates suggest that as borrowing costs increase, financial pressures on individuals rise, leading to a higher incidence of external fraud. Coming to EPWS, our results are aligned with those in Hambuckers *et al.* (2018) on the impact of the housing price index (HPI), but also find leverage (leveragelag) and log returns of UniCredit's value (logreturns) important. These variables reflect a bank's investment policy and market perception, which influence employment litigation outcomes. Moreover, GDP is not a significant driver unlike other loss categories, differing from Cope *et al.* (2012). Concerning internal frauds (IFRAUD), our findings show GDP (both EU and IT), S&P 500, M1, and HPI as important. This suggests that worsening economic and financial conditions increase the losses due to internal fraud. We remark that these findings contrast with those of Paredes and Vega (2024), who find the GDP not significant for the internal fraud losses. We do not observe deposit growth as an important variable, which differs from the findings of Povel *et al.* (2007).

6. Conclusion

Understanding the role of covariates for extreme losses is essential for proper risk management. This information can be provided by importance measures used in sensitivity analysis, but their estimation might be unstable in presence of small sample sizes. To overcome this issue, we increased the sample simulating new observations from the vine copula, to keep unaltered the original dependence structure in the observed data. We applied the proposed methodology to the operational losses of the UniCredit Bank, providing insights on the most important drivers of each type of loss, using Shapley effects as importance measure. Our analysis can be an important element supporting the quantitative analysis of operational risks. In particular, our method can be used to interpret machine learning models applied to operational loss data.

From a methodological perspective, we identify open research directions. First, we believe that the literature lacks a thorough study of the asymptotic behavior of Shapley effects estimated according to Mase *et al.* (2019). In this regard, there exists a central limit theorem for the permutation-based representation of the Shapley values (Benoumechiara and

Elie-Dit-Cosaque 2019), which, however, is an approximation of the true values and does not apply in our case, where we employed the exact representation. Having an asymptotic error bound could also help in determining the optimal augmented sample size to achieve a desired accuracy, without over-increasing the computational time. On this line, there are no results that link the error to the model's nonlinearity or to the distribution of the variables, an aspect that remains unexplored. Second, another area for future research is the exploration of the behavior of Shapley effects in the presence of highly dependent variables, as often encountered in financial and econometric datasets. In such a case, one possible approach could be to divide the variables into three groups—financial, macroeconomic, and firm-specific—and Principal Component Analysis (PCA) can be applied separately to each group, as done in Packham (2024) under the term *clustered PCA*. The Shapley effects can then be used on the resulting principal components to assess the relative importance of the three groups.

Finally, as mentioned above, we remark that our approach is not only useful for Shapley effects but can also be applied to other sensitivity measures when few extreme values are available, such as tail-distribution methods.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

The dataset has been downloaded from <https://doi.org/10.1002/jae.2638>.

References

- Abdymomunov, A., Curti, F. and Mihov, A., U.S. banking sector operational losses and the macroeconomic environment. *J. Money, Credit Banking*, 2020, **52**(1), 115–144.
- Bedford, T. and Cooke, R.M., Vines—a new graphical model for dependent random variables. *Ann. Stat.*, 2002, **30**(4), 1031–1068.
- Bee, M., Hambuckers, J. and Trapin, L., Estimating value-at-risk for the g-and-h distribution: An indirect inference approach. *Quant. Finance*, 2019, **19**(8), 1255–1266.
- Bee, M., Hambuckers, J. and Trapin, L., Estimating large losses in insurance analytics and operational risk using the g-and-h distribution. *Quant. Finance*, 2021, **21**(7), 1207–1221.
- Benoumechiara, N. and Elie-Dit-Cosaque, K., Shapley effects for sensitivity analysis with dependent inputs: Bootstrap and kriging-based algorithms. *ESAIM: ProcS*, 2019, **65**, 266–293.
- Bolker, B.M., Grasselli, M.R. and Holmes, E., Short communication: Sensitivity analysis of an integrated climate-economic model. *SIAM J. Financial Math.*, 2021, **12**(2), SC44–SC57.
- Borgonovo, E., A new uncertainty importance measure. *Reliab. Eng. Syst. Saf.*, 2007, **92**(6), 771–784.
- Borgonovo, E., Clemente, G.P. and Rabitti, G., Why insurance regulators need to require sensitivity settings of internal models for their approval. *Finance Res. Lett.*, 2024, **60**, 104859.
- Borgonovo, E., Figalli, A., Plischke, E. and Savaré, G., Global sensitivity analysis via optimal transport. *Manage. Sci.*, 2024.

- Borgonovo, E. and Plischke, E., Sensitivity analysis: A review of recent advances. *Eur. J. Oper. Res.*, 2016, **248**(3), 869–887.
- Borgonovo, E., Plischke, E. and Rabitti, G., The many Shapley values for explainable artificial intelligence: A sensitivity analysis perspective. *Eur. J. Oper. Res.*, 2024, **318**(3), 911–926.
- Brechmann, E.C. and Czado, C., Risk management with high-dimensional vine copulas: An analysis of the Euro Stoxx 50. *Stat. Risk Model.*, 2013, **30**(4), 307–342.
- Chavez-Demoulin, V., Embrechts, P. and Hofert, M., An extreme value approach for modeling operational risk losses depending on covariates. *J. Risk Insurance*, 2016, **83**(3), 735–776.
- Chernobai, A.S., Jorion, P. and Yu, F., The determinants of operational risk in U.S. financial institutions. *J. Financial Quant. Anal.*, 2011, **46**(6), 1683–1725.
- Cope, E.W., Piche, M.T. and Walter, J.S., Macroeconomic determinants of operational loss severity. *J. Bank. Financ.*, 2012, **36**(5), 1362–1380.
- Czado, C., *Analyzing Dependent Data with Vine Copulas: A Practical Guide with R*, Vol. 222, 2019 (Springer: Cham).
- Czado, C. and Nagler, T., Vine copula based modeling. *Annu. Rev. Stat. Appl.*, 2022, **9**(1), 453–477.
- Gamboa, F., Klein, T. and Lagnoux, A., Sensitivity analysis based on Cramér–von Mises distance. *SIAM/ASA J. Uncertainty Quantif.*, 2018, **6**(2), 522–548.
- Hambuckers, J., Groll, A. and Kneib, T., Understanding the economic determinants of the severity of operational losses: A regularized generalized pareto regression approach. *J. Appl. Econom.*, 2018, **33**(6), 898–935.
- Il Idrissi, M., Chabridon, V. and Iooss, B., Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs. *Environ. Model. Softw.*, 2021, **143**, 105115.
- Iooss, B. and Prieur, C., Shapley effects for sensitivity analysis with correlated inputs: Comparisons with Sobol’ indices, numerical estimation and applications. *Int. J. Uncertain. Quantif.*, 2019, **9**(5), 493–514.
- Mase, M., Mase, M., Owen, A.B. and Seiler, B.B., Explaining black box decisions by Shapley cohort refinement. arXiv: Learning, 2019.
- McNeil, A.J., Frey, R. and Embrechts, P., *Quantitative Risk Management: Concepts, Techniques and Tools*, 2015 (Princeton University Press: Princeton).
- Owen, A.B., Sobol’ indices and Shapley value. *SIAM/ASA J. Uncertainty Quantif.*, 2014, **2**(1), 245–251.
- Owen, A.B. and Prieur, C., On Shapley value for measuring importance of dependent inputs. *SIAM/ASA J. Uncertainty Quantif.*, 2017, **5**(1), 986–1002.
- Packham, N., Risk factor aggregation and stress testing. *Quant. Finance*, 2024, **24**(9), 1327–1340.
- Paredes, R. and Vega, M., An internal fraud model for operational losses in retail banking. *Appl. Stoch. Models. Bus. Ind.*, 2024, **40**(1), 180–205.
- Povel, P., Singh, R. and Winton, A., Booms, busts, and fraud. *Rev. Financ. Stud.*, 2007, **20**(4), 1219–1254.
- Rabitti, G. and Tzougas, G., Accelerating the computation of Shapley effects for datasets with many observations. *Eur. Actuar. J.*, 2025, **54**(1), 1–14.
- Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S., *Global Sensitivity Analysis: The Primer*, 2008 (John Wiley & Sons, Ltd: Chichester).
- Shapley, L.S., A value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, pp. 307–318, 1953 (Princeton University Press: Princeton).
- Sobol, I., Sensitivity estimates for nonlinear mathematical models. *Math. Modell. Comput. Exp.*, 1993, **4**, 407–414.
- Song, E., Nelson, B.L. and Staum, J., Shapley effects for global sensitivity analysis: Theory and computation. *SIAM/ASA J. Uncertainty Quantif.*, 2016, **4**(1), 1060–1083.
- Stübinger, J., Mangold, B. and Krauss, C., Statistical arbitrage with vine copulas. *Quant. Finance*, 2018, **18**(11), 1831–1849.
- Vallarino, A., Rabitti, G. and Khorrami Chokami, A., Construction of rating systems using global sensitivity analysis: A numerical investigation. *ASTIN Bulletin*, 2023, **54**(1), 25–45.