2023

Modelling Multivariate Dependence among Cyber Risks

Oliveira Ann Gyamfua Darkwah
Illinois State University, oliveiradarkwah@gmail.com

MODELLING MULTIVARIATE DEPENDENCE AMONG CYBER RISKS

Oliveira Ann Gyamfua Darkwah

37 pages

Over the years, there have significant number of cyber-attacks in various sectors causing losses to those sectors. We study the multivariate dependence of the cyber risks among these sectors by developing an iterative algorithm using copulas to estimate cyber risks in each sector by measuring the interdependencies and intradependencies among the sectors. The prediction performance of the proposed algorithm shows that the algorithm is superior to other methods.

KEYWORDS: Copula; Data breach

MODELLING MULTIVARIATE DEPENDENCE AMONG CYBER RISKS

Oliveira Ann Gyamfua Darkwah

A Thesis Submitted in Partial
Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

Department of Mathematics

ILLINOIS STATE UNIVERSITY

2023

MODELLING MULTIVARIATE DEPENDENCE AMONG CYBER RISKS

Oliveira Ann Gyamfua Darkwah

COMMITTEE MEMBERS:

Professor Maochao Xu, Chair

Professor Pei Geng

ACKNOWLEDGMENTS

I would like to thank my supervisor, Professor Maochao Xu for his guidance, support, and patience during my research. I also thank Professor Pei Geng for impacting knowledge to me and serving on the committee.

I am grateful to Professor George Seelinger, Professor Gaywalee Yamskulna, Professor Papa Sissokho, Professor Fuxia Cheng, Professor Wenhua Zhao, and Professor Olcay Akman for their acts of service, knowledge, and guidance which have enriched my knowledge and expertise in mathematics and statistics.

I appreciate the Mathematics Department and Illinois State University for the academic resources and support which have aided my master's degree pursuit.

Lastly, I thank my family, and friends for their enormous support.

O.A.G.D.

CONTENTS

TABLES

FIGURES

CHAPTER I: INTRODUCTION AND MOTIVATION

Cyber risk has been prevalent over the years causing great losses to individuals, businesses, and nations. It is defined as the combination of the probability of an event occurring within the realm of an organization's information assets, computer and communication resources and the consequences of that event for an organization (BIS, 2016). According to the Federal Bureau of Investigations Report (2022), the FBI's Internet Crime Complaint Center (IC3) reported cybercrime losses totaling $27.6 billion from 2018 to 2022 with significant increase in the losses over the years, which is alarming since it is causing financial strain on individuals, organizations, and states. Among the various types of cyber risks including online threats, physical threats, insider threats and data breaches, the type of interest in this study is data breaches, which refers to identity theft (in other words, losing personal information to someone). Data breaches are particularly of interest because according to the National Council on Identity Theft Protection, the Federal Trade Commission have received 5.7 million total fraud and identity theft reports in the year 2023, of which identity theft reports were 1.4 million (25%). Furthermore, the Identity Theft Resource Center reports data breaches from 2016 to 2022 of which 10,617,801,240 victims were impacted in the United States.

In our current age, where some operations of some sectors require the involvement of other sectors, the possibility of a cyber-attack on one sector causing an attack on other sectors is high. Macaulay & Centre for International Governance [CIG] (2019) after assessing the interdependencies of ten critical infrastructure sectors (energy, communications & IT, finance, health care, food, water, transportation, safety, government, and manufacturing) discovered that most sectors had high potential vulnerability due to cyber threats from other critical infrastructure sectors, in the dependency matrix. The inbound and outbound dependence sum up to interdependence which is ranked from 1 to 10 where a lower number implies low interdependence, and a higher number implies high interdependence.

Since some of the interdependence values are greater than 5, it necessitates studying the interdependence among sectors.

Inbound dependencies (vulnerabilities)

Outbound dependencies (threats)

Critical Infrastructure sector	Energy	Communications & IT	Finance	Health Care	Food	Water	Transportation	Safety	Government	Manufacturing
Energy	9.37	3.63	2.48	3.88	2.06	3.08	4.25	3.23	3.36	3.24
Communications & IT	6.96	8.82	4.48	5.11	2.32	3.42	4.41	4.62	3.96	7.08
Finance	7.13	7.19	8.95	4.23	8.23	5.01	6.78	4.02	5.18	7.96
Health Care	4.12	2.43	2.99	8.25	1.8	4.43	3.33	5.78	5.06	2.57
Food	1.47	1.66	1.94	3.76	6.45	1.83	2.48	1.05	2.71	1.99
Water	4.90	1.84	1.96	3.6	1.3	5.78	3.18	1.20	2.87	2.16
Transportation	6.82	3.95	4.23	4.95	5.06	2.96	7.49	3.78	4.66	5.84
Safety	7.85	3.96	3.6	5.71	1.02	4.54	5.35	8.23	5.73	4.96
Government	5.85	5.05	7	6.12	4.76	5.05	7.61	6.43	8.78	5.96
Manufacturing	5.87	3.75	4.66	5.01	4.5	3.43	4.53	1.17	3.63	7.15

Source: Macaulay & CIG (2019)
Figure 1: Dependency Matrix for 10 Critical Infrastructure Sectors

Therefore, we study the multivariate dependence among 14 sectors: education, financial services, government, health care, hospitality, manufacturing & utilities, military, non-profit organizations, professional services, retail, technology, transportation, other sectors, and uncategorized sectors. We combined military and government as one sector, as well as other and uncategorized sector as one sector. We develop a copula model using the vine copula approach to predict cyber risk in each sector given the others.

There are few literatures on modelling multivariate dependencies among cyber risks. Zhang et al. (2021) proposed a hybrid model which first models the multivariate attack time series by a deep learning model and then uses the extreme value theory to model residuals exhibiting heavy tails. Xu et al. (2017) also modelled the dependence among the time series of the number of cyber-attacks, and the dependence between the time series of the number of attacked computers using a vine copula

approach. Peng et al. (2018) also developed a copula – Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model which uses vine copulas to model the multivariate dependence among cyber-attacks. In addition to Xu et al (2017) and Peng et al (2018) who developed models using the vine copula approach, another literature on vine copulas is by Chang & Joe (2019) where they proposed a vine copula regression method to compute the conditional distribution of the response variable given the explanatory variables. To the best of our knowledge, there is no literature on the use of vine copula to model the multivariate dependence among cyber risks in the respective sectors.

In our study, we develop a new copula-based algorithm which is a build-up on the Additive Decomposition Algorithm Based on Copulas (ADABOC): an algorithm in which the response variable is decomposed as a sum of error terms. Each error term is estimated, identifying the input variable that best "copulates" with the error term (Carrillo et al., 2021). The ADABOC has two drawbacks: it is time-consuming and tends to select the same copula family and predictor variable in most iterations. Thus, we resolve those drawbacks by

i) creating a new simulation algorithm by generating samples from the conditional copula to approximate the average and
ii) using the idea of "decorrelation" to ensure that the algorithm can choose different copulas.

The authors stated that ADABOC is not oriented for time series data, however, the new copula-based algorithm we have developed using the ADABOC is superior to other well-known multivariate time series models.

The rest of the paper is presented as follows: we discuss the exploratory data analysis in Chapter II, highlight preliminaries of the proposed algorithm and other modelling methods in Chapter III, introduce the proposed algorithm in Chapter IV, and then, apply it on the dataset and compare its prediction performance with other models in Chapter V, and finally conclude in Chapter VI.

CHAPTER II: EXPLORATORY DATA ANALYSIS

The data represents reported data breach records published from March 25, 2020, to December 22, 2022, by the Identity Theft Resource Center. Although the publications are from 2020 to 2022, the dates of breach are from 1970 to 2022. However, for this study, we focus on the dates the data breaches were reported, that is, from 2017 to 2022, exempting 2016 due to many missing values in 2016. Government and military sectors were merged, as well as Other and uncategorized sectors due to many missing values for the military and uncategorized sectors.

Each sector's mean of impacted victims in Table 1 is greater than its median, signifying that each distribution of the impacted victims in each sector is positively skewed. The positive skewness of the sectors is also displayed in the boxplots in Figure 2. Technology has the highest mean number of impacted victims (53,077,771.95) followed by Other/uncategorized sectors (38,518,179.93). Non-profit organizations on the other hand, has the lowest mean number of impacted victims (95,215). Each number of NA implies no data breach was recorded for those number of months. Transportation sector had 19 NAs signifying that there were no reported breaches in 19 months out of the 72 months studies. The low mean number of impacted victims for Non-profit/NGO is accounted for by the 15 NAs the sector had. Technology had 7 NAs and yet had the highest mean since any time the data breach occurred, the number of victims impacted were many, whereas Other/uncategorized sectors had no NA which means the several occurrences of the data breaches contributed to the high mean. Similarly, education, financial services, and health care services have no NAs, thus, these sectors have data breach reports every month, which implies these sectors could be more prone to cyber-attacks. The manufacturing and utilities sector also had only one NA signifying they had data breaches in all the other 71 months.

Sector	Min.	Q1	Median	Mean	Q3	Max.	NA's
Education	150	14535	49136	873140	207044	39724935	0
Financial Services	145	75577	161638	2831605	829305	100090348	0
Health care	40909	381382	1049780	1915167	2208173	21547765	0
Hospitality	0	2566	22698	9452143	107903	383021054	16
Manufacturing & Utilities	2621	36673	67563	2646888	294639	48156106	1
Non-profit/NGO	192	4861	22183	95215	112569	1010676	15
Professional Services	703	14022	64310	1552608	310704	49030228	5
Retail	390	34978	138383	6520800	843271	218000412	2
Technology	625	66667	372052	53077771.95	8587705	1032936604	7
Transportation	0	3700	18094	1316932	81900	47600912	19
Other/Uncategorized sectors	3465	245020.75	775359.5	38518179.93	4636654.75	1383360666	0
Gov't/Military	0	11224	45626	1358173	254234	60009709	2

Table 1: Summary Statistics of the Various Sectors where Q1 and Q3 represent first and third quartiles respectively, Min. and Max. represent the minimum and maximum values respectively, and NA's represent missing values.

Also, from the boxplots in Figure 2, there are quite a high number of extreme values for technology and other/uncategorized sectors accounting for their high means. All sectors have small variability from the boxplots in Figure 2. In the data analysis, all NAs are considered as zeros.

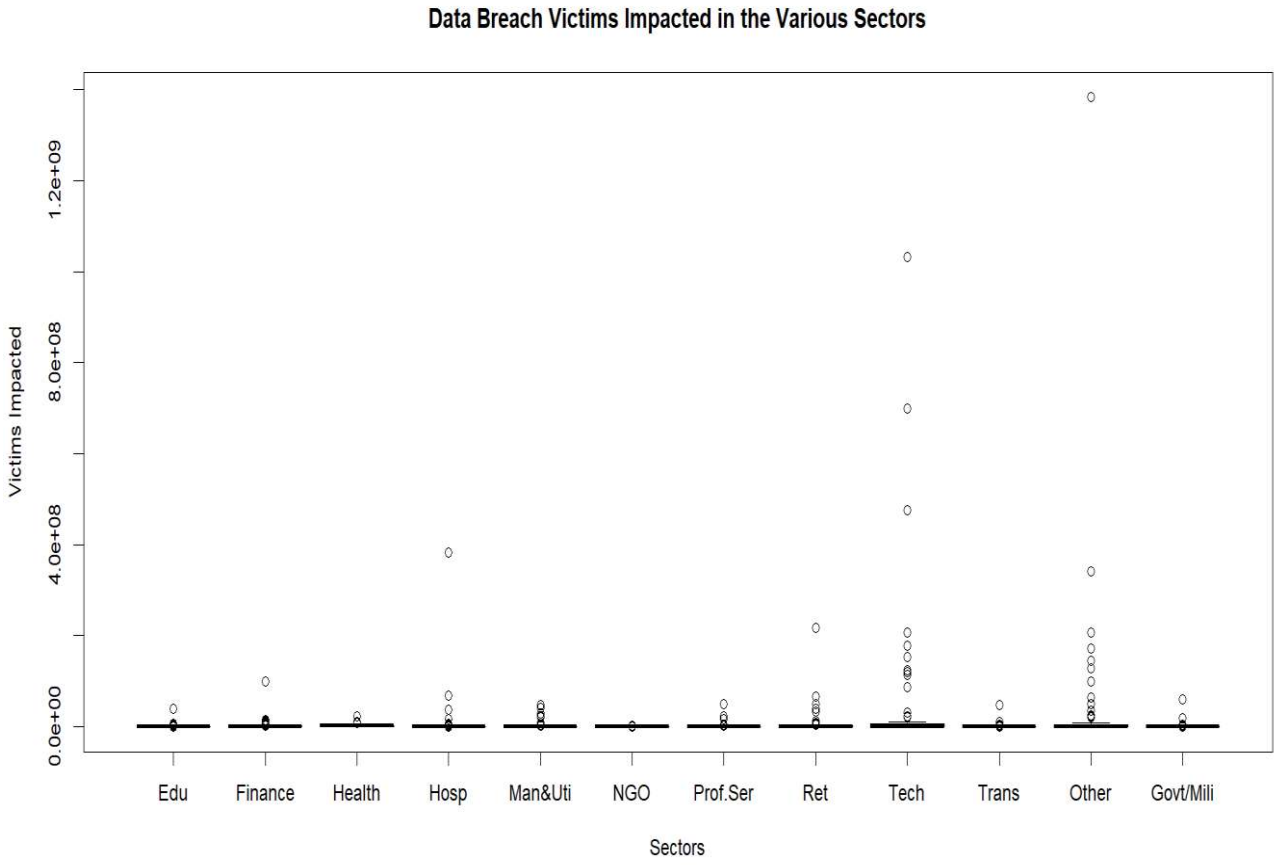
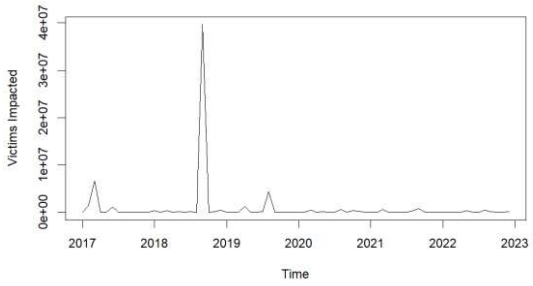
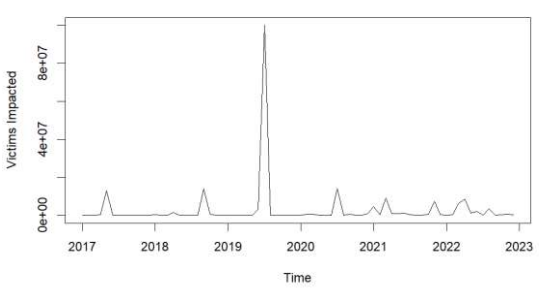


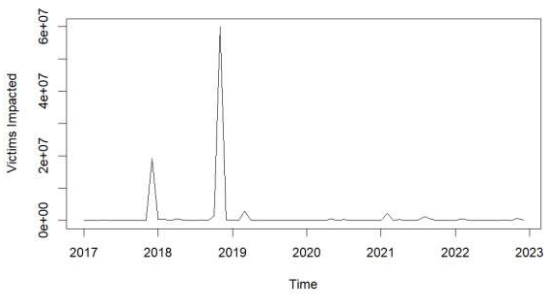
Figure 2: Boxplot on the Impacted Victims across the Various Sectors



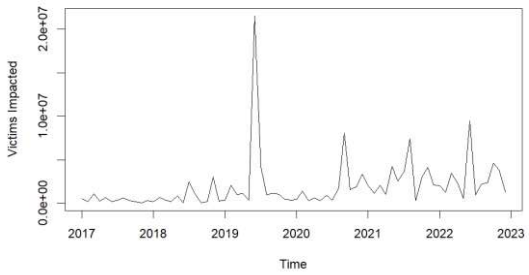
a) Education



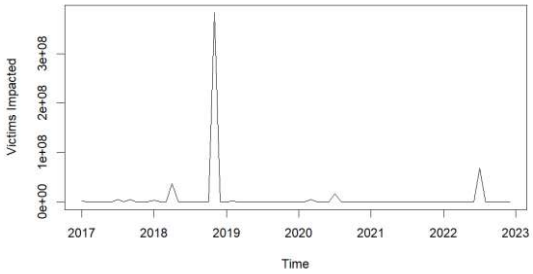
b) Financial Services



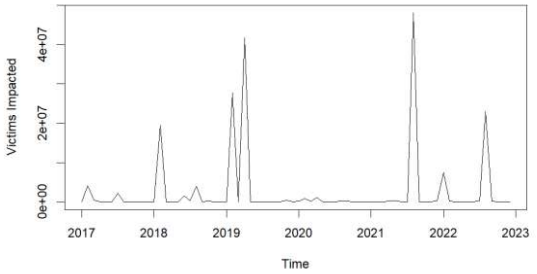
c) Government/military



d) Health care



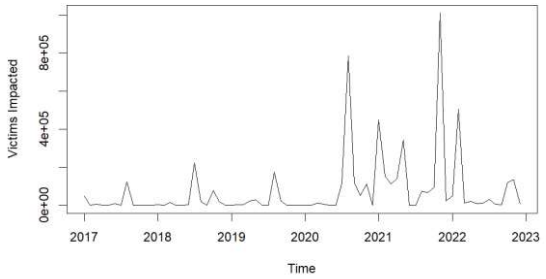
e) Hospitality



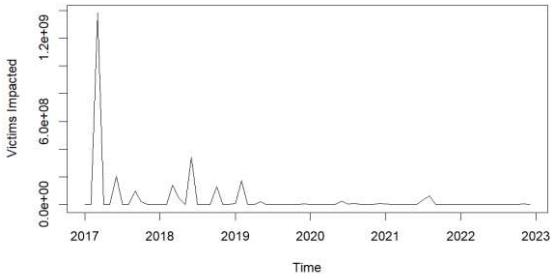
f) Manufacturing & Utilities

(Figure Continues)

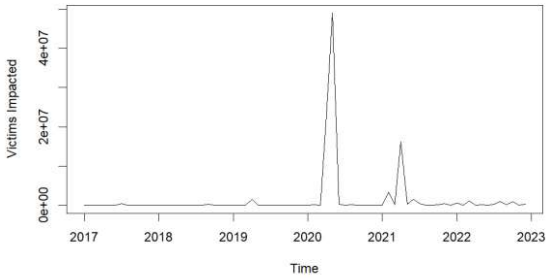
(Figure Continued)



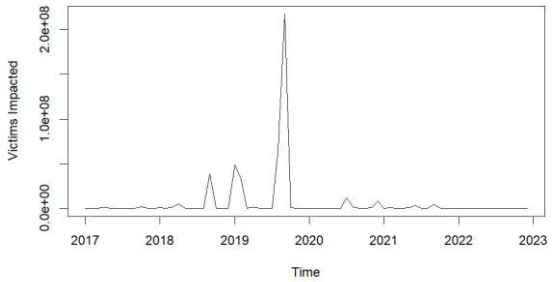
g) Non-profit / NGO



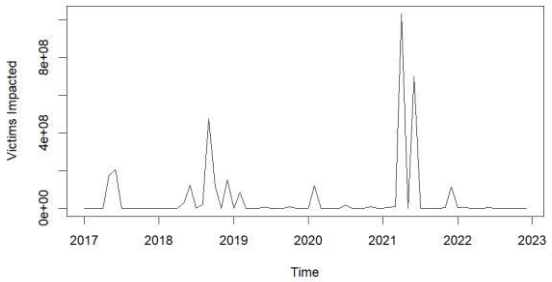
h) Other/uncategorized



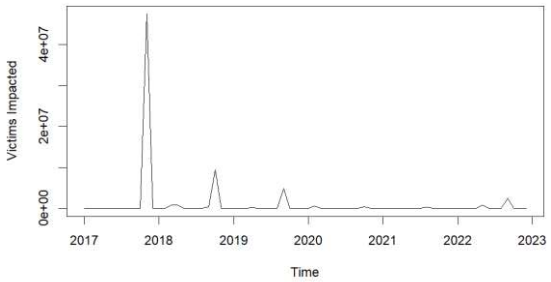
i) Professional Services



j) Retail



k) Technology



l) Transportation

Figure 3: Time Series Plots of the Various Sectors' Impacted Victims

From Figure 3, other/uncategorized and transportation sectors had a spike in 2017, manufacturing and utilities sector had a little spike in the early part of 2018, whereas education, hospitality, technology, and government/military had a spike towards the end of 2018. Financial services, retail, manufacturing and utilities, and health care also had a spike in 2019. NGO and healthcare had high numbers of impacted victims from 2020, whereas professional services had a spike in 2020. Most sectors have a reducing number of impacted victims from 2021 to 2022 except NGO. Although manufacturing and utilities sector also had a high number of impacted victims in 2021, the sector had a reduced number of impacted victims towards the end of 2021 through 2022. Furthermore, sectors including education, financial services, government, other, retail, and transportation have maintained a decreasing trend since the months they had large number of data breaches. Technology rather has more rise and falls which is of no surprise since cyber-attacks are related to computers, information, and technology.

CHAPTER III: PRELIMINARIES

We introduce some preliminaries used in this study.

III.1 Copula

Nelsen (2006) as cited in Chen & Guo (2019) defines copulas as functions that join or "couple" multivariate distribution functions to their one-dimensional marginal distribution functions. A precise definition is given in Sklar's theorem (Sklar, 1959, as cited in Chen & Guo, 2019) where he first introduced copulas in the statistical context:

If random variables X_1, ..., X_n follow an arbitrary marginal distribution function $F_1(x_1)$, ..., $F_n(x_n)$, respectively, there then exists a copula, C, that combines these marginal distribution functions to give the joint distribution function, $F(x_1, \dots, x_n)$ as follows:

$$F(x_1, \dots, x_n) = C\{F_1(x_1), \dots, F_n(x_n)\}, \qquad x_1, \dots, x_n \in R \qquad \textit{(Eqn. 1)}$$

If the marginal distributions $F_i(x_i)$ are continuous, the copula function C is unique.

Since X_i is a random variable with cumulative distribution function $F_i(x_i) = P(X_i \leq x_i)$, then, $F_i(x_i)$ is also a random variable (Carrillo et al, 2021). Thus, if the inverse of F_i exists, then,

$$P(F_i(X_i) \leq x_i) = P\left(X_i \leq F_i^{-1}(x_i)\right) = F_i\left(F_i^{-1}(x_i)\right) = x_i \qquad \textit{(Eqn. 2)}$$

Given a random variable $M \sim U(0,1)$, the cumulative distribution (K_M) of the standard uniform random variable is $K_M(m) = \frac{m-0}{1-0} = m$. Thus, it suffices to say that F_i is in accordance with the cumulative distribution of a standard uniform distribution, from Eqn. 2. This makes it appropriate to consider $F_1(x_1)$, …, $F_n(x_n)$ as uniform variables $U_1, \dots, U_n$ respectively. Therefore, Eqn. 1 can be rewritten as:

$$F(x_1, \dots, x_n) = C\{F_1(x_1), \dots, F_n(x_n)\} = C(u_1, \dots, u_n) \qquad \textit{(Eqn. 3)}$$

Bivariate copulas are used in this case since we are modelling the dependence between pairs of sectors. A bicopula is a function $C: [0,1] \times [0,1] \rightarrow [0,1]$ with the following properties (Carrillo et al, 2021):

1. $C(u, 0) = 0; C(0, w) = 0;$
2. $C(u, 1) = u; C(1, w) = w;$
3. *C is non-decreasing for each hyperrectangle* $B = [u_1, u_2] \times [w_1, w_2]$*, its volume is non-negative:*

$$V_C(B) = C(u_2, w_2) - C(u_2, w_1) - C(u_1, w_2) + C(u_1, w_1) \geq 0$$

where (u, w) represent the uniform pairs of the two random variables.

Conditioning with copulas

To make predictions of a variable Y given another variable, X, we can calculate the expected values of Y by finding the conditional expectation of Y given X. Calculating the conditional expectation involves the use of the conditional density function of Y|X, which is computed from the joint density function of X and Y, and the marginal distribution of X. In our case, since a bicopula function will be the joint density function, it is important to introduce conditional copulas to enable us to compute the conditional expectation of Y|X. Venter (2001) adapted Sklar's theorem to continuous conditioned distribution in the following proposition:

Let C be a copula function and let $C_1(u, w)$ denote the derivative of C(u, w) with respect to the first argument. When the joint distribution of X and Y is given by $L(x, y) = C(F_X(x), Q_Y(y))$*, then the conditional distribution of Y|X=x is given by:*

$$Q_{Y|X}(y) = C_1(F_X(x), Q_Y(y)).$$

Carrillo et al (2021) presented the definition of conditioned copulas associated with a copula as:

Fixed $U = u$*, copula conditioned to* u *is a function on W variable:* $C_1(u, w) = C(w|u) = \frac{dC(u,w)}{du}$.

Fixed $W = w$*, copula conditioned to* w *is a function on U variable:* $C_2(u, w) = C(u|w) = \frac{dC(u,w)}{dw}$.

Thus, given the product/independent copula $C(u, w) = u.w$, the conditional distribution of U given W is

$$C_2(u, w) = C(u|w) = \frac{dC(u, w)}{dw} = \frac{d\prod(u, w)}{dw} = \frac{d(u.w)}{dw} = u$$

which implies U is independent of the value of W.

III.2 Vector Autoregressive (VAR) Model

We compare the proposed algorithm with a time series model (the Vector Autoregressive model). Therefore, we define it as follows:

According to Tsay (2005) a time series $\boldsymbol{r}_t$ follows a VAR (p) model if it satisfies

$$\boldsymbol{r}_t = \emptyset_0 + \Phi_1 r_{t-1} + \cdots + \Phi_p r_{t-p} + \boldsymbol{a}_t, \quad p > 0$$

where $\emptyset_0$ is a k-dimensional vector, Φ_j are $k \times k$ matrices for $j = 1, \ldots, p$, and $\{\boldsymbol{a}_t\}$is a sequence of serially uncorrelated random vectors with mean zero and covariance matrix Σ which is required to be positive definite in application.

III.3 Regression methods

Due to the nature of the data, that is, *multivariate* time series data, we also compare the proposed algorithm with some regression methods: least squares regression, ridge regression, lasso regression, and two ensembled trees methods (random forest and extreme gradient boosting). While the least regression simply measures the linear relationship between predictor variables and the dependent variables, the ridge regression can shrink the estimated coefficients towards zero using the shrinkage penalty $\lambda(\sum_{j=1}^{p} {\beta_j}^2)$ where $\lambda \geq 0$ is a tuning parameter, to minimize the sum of squares

residuals (James et al, 2013). Like the ridge regression, the lasso regression minimizes the sum of squares residuals by forcing some estimated coefficients to zero using the shrinkage penalty $\lambda(\sum_{j=1}^{p}|\beta_j|)$ where $\lambda \geq 0$ is a tuning parameter. The random forest and extreme gradient boosting have more accurate predictions among these five methods, in that, the random forest uses feature randomness, that is, it generates a subset of the predictor variables to ensure low correlation among the decision trees; the average of all the decision trees prediction is the final prediction (James et al, 2013). The extreme gradient boosting (XGBoost) also uses multiple trees by iteratively training an ensemble of shallow decision trees; in each iteration, we fit the next model with the error residuals of the previous model and make the final prediction using the weighted sum of the decision tree predictions (NVIDIA, 2023).

III.4 Accuracy metrics

We use the Mean Absolute Error (MAE) to evaluate the prediction accuracy of the various models. MAE is represented as $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ where n is the number of observations, y_i is the values of the response variable and $\hat{y}_i$ is the estimated values of the response variable.

CHAPTER IV: PREDICTION ALGORITHM BASED ON BICOPULA FUNCTIONS

We study the prediction algorithm involving bicopula functions. The data is grouped into training, validation, and test datasets. The training dataset is used to train the model, primarily to determine the initial prediction of the target variable, and the copula and predictor variable to select in each iteration. The validation dataset contributes to the termination of the iterations. Like the extreme gradient boosting approach where the test root mean square error determines the best number of rounds, that is, the number of rounds at which the minimum test root mean square is obtained (beyond this round, the test root mean square increases), the minimum validation MAE also determines the best number of iterations which would result in the best prediction. Lastly, the test dataset set is used to measure the prediction accuracy of the copula model.

In this algorithm, we first predict the response variable for each of the training, validation, and test datasets as the mean of the response variable of the training dataset. The first error values $\varepsilon_1(i)\ \forall\ i = 1, \dots, n_1$ of the training dataset are obtained by finding the difference between the observed values of the response variable and the predicted values of the response variable ($\overline{Y}_{train}$). Since copula functions join multivariate distribution functions to the unidimensional marginal distribution functions (Nelsen, 2006, as cited in Chen & Guo, 2019), we need to find the cumulative distribution function of each predictor variable as well as the errors. Thus, we estimate the cumulative distribution $F_j(X_j) = P(X_j \leq x_j) = U_j$ for all $j = 1, \dots, p$ in the case of each dataset and the error values obtained in the training dataset ($Q_k(\varepsilon_k)) = W_k$ using empirical estimators (DeMatteis, 2001). Using the vine copula package in R, we select the copula which best explains/models the relationship between the predictor variables and the error term (X_j, ε_k) of the training dataset in each iteration k. Given the 40 bicopula families $\{(C^1, \dots, C^a)\ \forall\ a\ \epsilon\ 1, \dots, 40\}$ and p predictor variables, we end up with $40p$ bicopula functions, one per predictor variable and copula family (Carrillo et al, 2021) in each iteration:

$$L_{jk}(X_j, \varepsilon_k) = C_{jk}\left(F_j(X_j), Q_k(\varepsilon_k)\right) = C_{jk}(U_j, W_k)$$

The best copula and pair ${C_{jk}}^*(U_j^*, W_k)$ with the minimum AIC are selected (Schepsmeier et al, 2015). We simulate *N* error values using the conditional copula simulation on U_j^* (Chang & Joe, 2019). We then use the quantile function to find the inverse of all simulated values (Hyndman & Fan, 1996). At this point, we find the average (either mean or median) of these values to estimate the respective error terms. In the case where the mean of the target variable is significantly lesser than the first quartile, or the other way — significantly greater than the third quartile, then, finding the median of the simulated values is preferred since the median is not affected by outliers. Otherwise, the mean is used in finding the average of the simulated values, that is, $\hat{\varepsilon}_{i,k} = \frac{1}{N}\sum_{b=1}^{N} e_b(i)\ \forall\ i \in 1, \dots, n$, where the e_i's represent the inverse of the simulated values. The first estimated error term $\hat{\varepsilon}_1$ is added to the initial prediction of the response variable ($\bar{Y}$) to estimate the response variable. Note that the initial predictor is the mean of the target variable, therefore, extreme values are accounted for in the overall prediction even when we use the median as the average of the simulated values. This step is repeated in the validation and test datasets using the same copula and the predictor variable used in the case of the training dataset. Thus, the variables in the training dataset should be the same variables in the validation and test datasets.

In the next iterations, we repeat similar steps, only that

i. The residuals ε_k will be the difference between the observed target values and the estimated target values from the previous iteration. We estimate the cumulative distribution function of the residuals generated in each iteration using empirical estimators to ensure the residuals are in the copula domain [0,1].

ii. We sample 50% of the predictor variables during the copula selection stage. Here, we ensure the algorithm can choose different copulas in the various iterations by using the idea of "decorrelation". Otherwise, there is high probability of selecting the same copula and predictor variable in each iteration. This sampling approach is similar to the random forest method where a random sample of *m* predictors are considered as split candidates from the full set of *p* predictors when building a decision tree (James et al, 2013). After several simulation studies using 30% to 90% sample of the predictor variables, 50% sample had the best prediction accuracy, which explains why we sample 50% of the predictor variables.

The algorithm stops if

i. The selected copula is the independence / product copula since that will indicate independence between the predictor variable and the error variable. (Carrillo et al, 2021) refers to this as the *independence criterion*.

ii. The validation MAEs of the last 10 iterations are all greater than that of the $(k-10)^{th}$ iteration. This is also referred to as the *early stopping criterion*.

iii. The maximum number of iterations specified are reached.

Thus, the estimated target variable is

$$\hat{Y}_{train} = \overline{Y}_{train} + \sum_{k=1}^{k^*} \hat{\varepsilon}_{k_{(train)}}$$

$$\hat{Y}_{valid} = \overline{Y}_{train} + \sum_{k=1}^{k^*} \hat{\varepsilon}_{k_{(valid)}}$$

$$\hat{Y}_{test} = \overline{Y}_{train} + \sum_{k=1}^{k^*} \hat{\varepsilon}_{k_{(test)}}$$

where k^* is the iteration with the lowest MAE in the validation data set.

Algorithm 1: Modified ADABOC

INPUTS:		
	Y	Target/response variable
	$(X_1, \dots, X_p)$	Independent variables
	$(y(i), x_1(i), \dots, x_p(i))$	Data observations $\forall\ i \in 1, \dots, n$
	n_1	Number of train data observations
	n_2	Number of validation data observations
	n_3	Number of test data observations
	N	conditional copula simulation number of values
	maxiter	maximum number of iterations
	average	median or mean
	$\{C^a\}$	Bicopula families $\forall\ a \in 1, \dots, 40$

counterNoImprovement Cumulative number of iterations without improvement counter (Carrillo et al, 2021) (set to 10 in this study)
$n = n_1 + n_2 + n_3$

1) Estimate the first predictor of Y, which is the first predictor for each dataset: $\hat{y}_0(i) = \frac{1}{n_1}\sum_{i=1}^{n_1} y(i)$.
2) At k=1, first error variable : $\varepsilon_1(i) = y(i) - \hat{y}_0(i), \forall\ i \in 1, \dots, n_1$
3) **for** k = 1, …, *maxiter*, **do**
4) Estimate the marginal distributions of each predictor variable in each dataset and error variable from the train data using empirical cumulative distribution function (DeMatteis, 2001):
   $F(X_j) = ecdf(x_j) = U_j \quad \forall\ j \in 1, \dots, p$
   $Q(\varepsilon_k) = ecdf\ (\varepsilon_k) = W_k$
5) $\left(u_j(i), w_k(i)\right) = \left\{F_{X_j}\left(x_j(i)\right), Q_{\varepsilon_k}\left(\varepsilon_k(i)\right)\right\} \forall\ i = 1, 2, \dots, n_1\ \forall\ j = 1, 2, \dots, p$ *(Pair in copula domain)*
6) **if** k = 1, **then** use all predictor variables
7) **else** sample 50% of the predictor variables.
8) Select copula which models the relationship between (U_j, W_k), with the minimum AIC: $(C_k{}^*, X_{jk}) = min_{j,a}\{AIC\left(C_{jk}{}^a\right)\}$ (Schepsmeier et al, 2015).
9) **if** $C_k{}^* = \prod$, *(independence criterion)*
10) **then** k = maxiter +1
11) **else, for** i = 1, …, n **do**
12) Simulate N values from copula $(C_k{}^*|U_j{}^*(i) = u_j{}^*(i))$.
13) Invert all simulated values using the quantile function to obtain $(e_{1k}(i), \dots, e_{Nk}(i))$ (Hyndman & Fan, 1996).
14) Compute the estimated error: $\hat{\varepsilon}_k(i)$= average $(e_{1k}(i), \dots, e_{Nk}(i))$.
15) Calculate the new Y estimator: $\hat{y}_k(i) = \hat{y}_{k-1}(i) + \hat{\varepsilon}_k(i)$
16) **end for**
17) Calculate the new error variable: $\varepsilon_{k+1}(i) = y(i) - \hat{y}_k(i)\ \forall\ i = 1, \dots, n_1$
18) **end if**
19) Compute $MAE_k = \frac{1}{n_2}\sum_{i=1}^{n_2} |y(i) - \hat{y}_k(i)|$ on validation dataset.
20) **if** $MAE_k < MAE_{k-1}$, **then** *counterNoimprovement* = 0,

21) $k^* = k$ *(Index associated with the iteration with minimum MAE)*
22) **else** *counterNoImprovement* = *counterNoImprovement* (=10 in this study) *(early stopping criterion)*
23) **end if**
24) **end if**
25) **end for**
26) Copulamodel = $\{\hat{Y}_0 = \bar{Y}_{train}, (X_{j1}, {C_1}^*, \varepsilon_1), (X_{j2}, {C_2}^*, \varepsilon_2), \ldots, (X_{jk^*}, {C_{k^*}}^*, \varepsilon_{k^*})\}$
27) **return** Copulamodel

CHAPTER V: STATISTICAL MODELLING

In this section, we develop models using the modified ADABOC, the ADABOC, VAR, and the five regression methods used in this study, and compare the prediction performance of the modified ADABOC with the other models.

V.1 Model Fitting

The ITRC data had 72 months of data breach reports from 2017 to 2022. Therefore, 60 months, that is, 2017 to 2021 were used to train the model, January 2022 to June 2022 formed the validation dataset and July 2022 to December 2022 formed the test data. Macaulay & CIG (2019) emphasized the interdependencies and intradependencies among the sectors. Thus, in order to study the interdependencies among the sectors, we select one sector as the target variable and the other sectors as the predictor variables. To address the intradependencies, we predict the response variable using r lags of the same variable. Therefore, the dataset will be of the form:

Target variable	**Predictor variables**					
Target sector (r+1)	Target sector (r)	⋯	Target sector (1)	Sector 1 (r)	Sector 2 (r)	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Target sector (72)	Target sector (71)	⋯	Target sector (72-r)	Sector 1 (71)	Sector 2 (71)	⋯

Table 2: Data Structure using r lags

Thus, Target sector (r) in Table 2 for example, implies the r^{th} observation in the target variable. To determine the lag r which optimizes the results, we used cross validation on the validation dataset.

Sector (Target variable)	**Lag 1**	**Lag 2**	**Lag 3**	**Lag 4**	**Lag 5**
Education	138,309.15	158,280.67	151,914.67	*137,051.41*	146,464.83
Financial Services	2,780,650.36	2,836,927.35	2,706,592.33	2,270,262.04	*2,174,910.52*
Health care	1,957,530.68	1,962,149.92	1,959,331.21	*1,708,517.01*	1,922,925.33
Hospitality	*15,078.48*	17,119.45	19,221.72	20,397.90	23,342.36
Manufacturing & Utilities	1,283,089.87	1,286,367.91	1,316,136.67	1,385,488.61	*1,278,348.56*
Non-profit/NGO	104,950.00	108,466.00	*103,766.13*	108,073.55	107,026.48
Professional Services	505,716.88	*376,626.91*	435,325.26	495,152.61	377,929.70
Retail	156,798.01	168,920.32	197,711.48	108,657.46	*108,037.88*
Technology	2,482,247.51	2,508,706.18	2,376,702.11	*2,112,416.76*	2,640,665.35
Transportation	*136,263.39*	137,331.00	136,857.47	144,383.94	149736.67
Other/Uncategorized sectors	1,247,677.26	1,259,639.23	1,131,931.05	*1,046,072.10*	1,212,020.96
Gov't/Military	*1,007,080.84*	1,601,545.30	1,303,834.29	1,034,414.52	1,387,551.14

Table 3: Cross validation results using maximum of five lags.

The cross-validation results from Table 3 show hospitality, transportation, and government/military have optimal results at lag 1, while professional services have optimal results at lag 2. NGO on the other hand has lag 3 as its optimal lag. Lag 4 gives optimal results for education, health, technology and *other*, while lag 5 is the optimal lag for finance, manufacturing & utilities, and retail. These lags are considered optimal since the validation MAEs at those lags are the lowest in each case. This reformed the dataset when building models for each sector, in that, there were *optimal lag* additional predictor variables to the already existing predictor variables (the other sectors), i.e., an optimal lag of 4 for the education sector reformed the dataset by including 4 more predictor variables (Education lags 1 to 4) to the other sectors. These reformed datasets were used in developing the modified ADABOC models, as well as all the other models except the vector autoregressive model since it determines its own optimal lag. Using the *vars* package in R, lag 5 was the selected lag using the AIC, the Schwarz Criterion (SC), and the Hannan-Quinn criterion (HQ), while lag 4 was selected

using the Final Prediction Error (FPE) criterion. Although three of the criteria selected lag 5, lag 4 was used in building the VAR model since VAR(5) forecasted NA's during the prediction stage while VAR (4) forecasted actual values.

Modified ADABOC					
Iteration	**Copula**	**Variable**	**Train MAE**	**Validation MAE**	**Test MAE**
1	Tawn type 2 copula	Transportation_l1	1,251,724	221,142	284,526
2	rotated Joe copula (180 degrees; survival Joe)	Professional Services_l1	1,253,755	224,995	273,187
3	rotated Tawn type 1 copula (270 degrees)	Non.Profit.NGO_l1	1,252,010	237,560	302,787
4	rotated Tawn type 2 copula (180 degrees)	Technology_l1	1,252,460	237,159	303,753
5	*Tawn type 2 copula*	*Manufacturing &Utilities_l1*	1,245,319	201,213	345,135
6	Tawn type 2 copula	Education_l1	1,224,281	441,070	547,833
7	rotated Tawn type 2 copula (270 degrees)	Professional Services_l1	1,225,666	419,614	544,632
8	rotated Joe copula (180 degrees; survival Joe)	Professional Services_l1	1,223,149	444,551	549,822
9	rotated Joe copula (180 degrees; survival Joe)	Healthcare_l1	1,221,414	444,272	547,103
10	rotated Tawn type 1 copula (180 degrees)	Hospitality_l1	1,221,100	451,802	554,650
11	rotated Tawn type 1 copula (180 degrees)	Education_l1	1,221,919	461,613	565,937
12	rotated Tawn type 2 copula (180 degrees)	Other_l1	1,220,267	467,906	564,840
13	rotated Tawn type 2 copula (180 degrees)	Manufacturing &Utilities_l1	1,219,568	471,407	568,261
14	rotated Tawn type 1 copula (270 degrees)	Non-Profit (NGO)_l1"	1,218,913	486,706	587,372
15	rotated Tawn type 2 copula (180 degrees)	Other_l1	1,218,766	486,741	588,459
ADABOC					
1	BB7	Transportation_l1	1,463,270	484,695	1,239,695
2	Rotated BB8 270 degrees	Transportation_l1	1,280,889	288,892	709,634
3	*Rotated BB8 270 degrees*	*Transportation_l1*	1,302,702	224,693	432,685
4	Survival BB8	Transportation_l1	1,414,414	534,334	972,614
5	Rotated BB8 270 degrees	Transportation_l1	1,283,160	305,997	562,691
6	Rotated BB6 270 degrees	Transportation_l1	1,298,420	226,443	350,662
7	Survival BB8	Transportation_l1	1,404,625	575,951	936,151
8	Rotated BB6 270 degrees	Transportation_l1	1,269,345	353,135	562,838
9	Survival BB8	Manufacturing & Utilities_l1	1,579,946	792,116	1,111,557

(Table Continues)

(Table Continued)

10	Rotated BB8 270 degrees	Manufacturing & Utilities_l1	1,310,895	346,995	626,160
11	Survival BB8	Other_l1	1,533,764	308,755	584,714
12	Rotated BB8 270 degrees	Other_l1	1,332,322	370,198	646,248
13	Rotated BB6 270 degrees	Transportation_l1	1,382,573	264,803	415,797
14	Survival BB8	Transportation_l1	1,410,709	564,876	992,726

Table 4: Copula models output of the modified ADABOC and ADABOC for government/military.

We display the output of both the modified ADABOC and the ADABOC models with the government/military sector as an example to portray the difference in the two algorithms, most importantly, to observe the improvement due to the modifications made on the ADABOC algorithm. From the output in Table 4, the modified ADABOC algorithm stops on the 15th iteration since the last ten validation MAEs were all greater than the validation MAE prior to them. The 5th iteration is preferred since it has the lowest validation MAE among all the iterations. Thus, the prediction for the government/military sector:

$$\hat{Y}_{gov't/military} = \hat{Y}_0 + \sum_{k=1}^{5} \hat{\varepsilon}_k.$$

The ADABOC on the other hand stopped on the 14th iteration since the 4th to 14th validation MAEs were all greater than that of the 3rd iteration which was the lowest. Thus, the predictor in this case is

$$\hat{Y}_{gov't/military} = \hat{Y}_0 + \sum_{k=1}^{3} \hat{\varepsilon}_k.$$

It is observed that the ADABOC algorithm selected the variable, Transportation_l1 (which implies Transportation lag 1) throughout the first 8 iterations before it selected the other two variables: Manufacturing & Utilities_l1 and Other_l1. Since Transportation_l1 most explains the residuals of the government/military sector, it was selected repeatedly before the other variables had the chance to be selected. This problem is solved in the modified ADABOC through the 50% sampling of the predictor variables. Thus, after selecting the variable which most explains the residuals in the first iteration which uses all the predictor variables, other variables were also selected through sampling. The percentage 50 keeps a balance in the variable selection such that it is not too high to increase the probability of the most correlated variable with the residuals to be selected in each iteration and it is not too low to cause only the least correlated variables with the residuals to be selected. Hence, in the modified ADABOC output for government/military, the variables: Professional services_l1, NGO_l1,

Technology_l1, and Manufacturing_l1 form part of the set of variables for making predictions on the government/military sector, in addition to the most correlated variable, Transportation_l1 since they were the variables selected in the best iterations (1 to 5), whereas only Transportation_l1 will be used in the case of the ADABOC since it was the only variable selected in the best iterations (1 to 3). Furthermore, the ADABOC algorithm tends to select almost the same copula through the iterations due to the repetitive selection of the same variable. While BB7 and Rotated BB8 270 degrees will be the only copulas used in the ADABOC prediction of the government/military sector, Tawn type 2 copula, rotated Joe copula (180 degrees; survival Joe), rotated Tawn type 1 copula (270 degrees), and rotated Tawn type 2 copula (180 degrees) will be used in the modified ADABOC prediction.

V.2 Prediction and Evaluation

Aside the prediction made on the validation and test datasets, the modified ADABOC model can predict target variables given a dataset of the same predictor variables (scoring data) as the training dataset using Algorithm 2. The algorithm uses the same copula and predictor variables selected in each iteration from the modified ADABOC model to simulate *N* values which is averaged (using the same average in the modified ADABOC model) and added successively to the initial predictor ($\bar{Y}_{train}$) up until the best iteration (k^*).

Algorithm 2: Modified ADABOC Prediction

INPUTS:	$\{\hat{Y}_0 = \bar{Y}_{train}, (X_{j1}, C_1{}^*, \varepsilon_1), \ldots, (X_{jk^*}, C_{k^*}{}^*, \varepsilon_{k^*})\}$	From modified ADABOC
	Scoring data	The dataset to make predictions on. It should only contain the same predictor variables as the training data.

1) Estimate the marginal distributions of each predictor variable in the scoring dataset using empirical cumulative distribution function:
   $F(X_j) = ecdf\ (X_j)_{score} = S_j$; for all $j \in 1, \ldots, p$
2) **for** i = 1, …, nrow(score data), **do**
3) $\hat{y}_0(i) = \bar{Y}_{train}$
4) **for** k = 1, …, k^***do**
5) Simulate *N* values from copula C_k conditioned on $S_{jk}(i)$.
6) Invert all simulated values using the quantile function to obtain $(e_{1k}(i), \ldots, e_{Nk}(i))$.
7) Compute the estimated error: $\hat{\varepsilon}_k$= average$(e_{1k}(i), \ldots, e_{Nk}(i))$
8) Find the estimated Y value: $\hat{y}_k(i) = \hat{y}_{k-1}(i) + \hat{\varepsilon}_k(i)$
9) **end for**
10) **end for**
11) **return** $\hat{y} = \{\hat{y}_{k^*}(1), \hat{y}_{k^*}(2), \ldots, \hat{y}_{k^*}(n_{score})\}$.

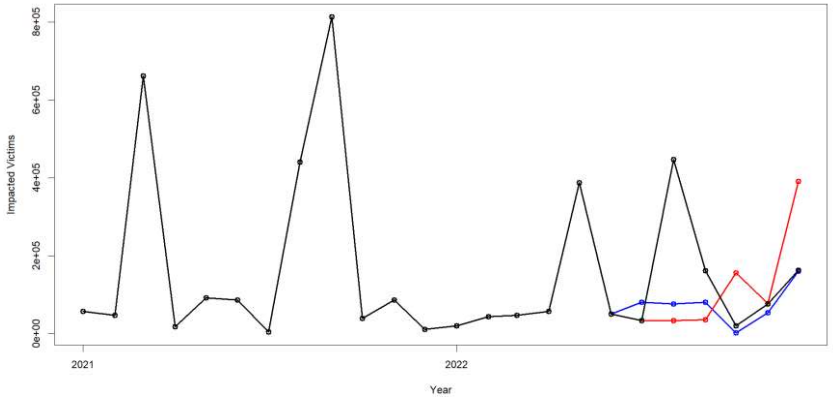
The prediction performance of the modified ADABOC is evaluated by comparing with the ADABOC, VAR and the other regression models. As mentioned earlier, the reformed datasets determined by the optimal lag in the case of each sector were used in building the ADABOC and the regression models while the VAR lag was 4 throughout all the sectors' model building and prediction. The least squares regression models for all the sectors had negative adjusted R^2 like the random forest (using 1000 trees) which also had negative percentage variability explained in all the sectors' models, signifying that these models do not explain the variability in the target sectors well. Using the minimum tuning parameters for the respective ridge regression and lasso regression models, the ridge regression shrunk all the coefficients very close to zero while the lasso regression turned most and, in some cases, all the coefficients to zero leaving only the intercept. In the case of the extreme gradient boosting, a maximum tree depth of 3 was used. We specified optimal number of rounds (beyond which the test root mean square error increased), after studying the performance of the model on the test

data in 70 rounds. Therefore, the XGBoost models for manufacturing & utilities, NGO, professional services, retail, technology, and *other* sectors used 1 round, education model used 2 rounds, government used 3 rounds, and health and hospitality used 4 and 6 rounds respectively. Sectors with higher number of rounds include finance (with 27) and transportation (with 66).

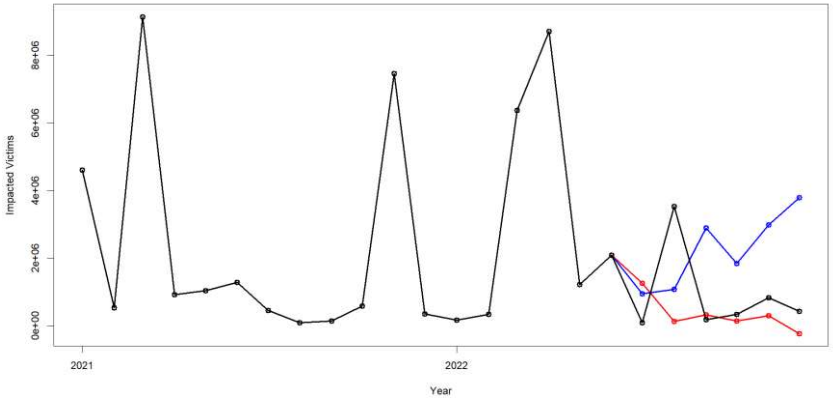
Sector (Target variable)	Modified ADABOC	ADABOC	Random forest	Linear model	Ridge regression	Lasso regression	VAR	XGBoost
Education	150,890	**89,967**	492,307	1,231,603	805,647	786,858	15,106,957	105,666
Financial Services	**1,015,175**	2,174,617	2,148,733	11,421,952	2,815,040	2,289,722	22,167,535	1,268,267
Health care	1,716,767	1,595,721	1,411,657	1,248,234	1,193,709	**1,170,167**	10,883,622	1,597,638
Hospitality	**11,556,906**	15,846,497	12,785,512	15,795,029	16,615,851	16,621,598	31,347,215	11,670,010
Manufacturing & Utilities	**3,916,881**	4,325,940	5,869,997	7,015,828	5,586,935	5,555,538	19,535,411	5,276,166
Non-profit/NGO	51,811	53,373	86,251	65,332	59,298	59,333	440,608	**46,486**
Professional Services	**514,520**	695,305	2,883,003	1,683,166	1,210,317	1,210,980	23,547,644	2,695,241
Retail	**71,159**	1,653,420	2,268,886	7,394,616	8,120,889	8,122,380	19,166,665	438,875
Technology	**1,574,423**	34,854,401	95,781,199	69,825,840	60,739,419	60,734,813	505,744,013	55,175,892
Transportation	**429,460**	552,613	464,017	1,261,471	1,156,639	1,108,753	6,792,692	472,876
Other/Uncategorized sectors	**1,317,992**	10,896,596	16,070,645	19,451,661	22,981,934	21,614,623	130,935,015	1,466,862
Gov't/Military	**345,135**	432,685	279,128	711,618	1,297,074	1,363,434	1,231,906	179,512

Table 5: Test MAE results by algorithm

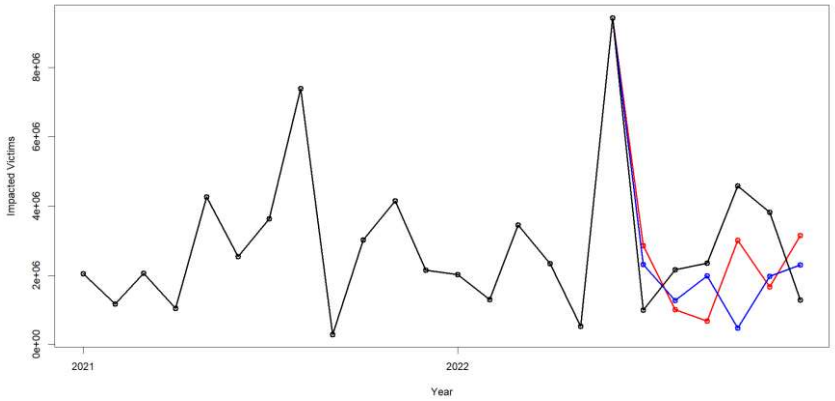
1-step ahead forecasts were made with the VAR(4) model twelve times to make forecasts for the entire year 2022 but the forecasted values for the months July 2022 to December 2022 were used in calculating the test MAEs since those months made up the test dataset. Therefore, the test MAEs for all the methods are displayed in Table 5. It is observed that the modified ADABOC had the lowest test MAE on nine of the sectors. Even in the case of education, healthcare, and NGO where the ADABOC, ridge regression and XGBoost had the lowest MAEs respectively, the modified ADABOC had the 3rd lowest, 11th lowest, and the 2nd lowest MAEs respectively. Therefore, the modified ADABOC has achieved competitive results. We also study the prediction performance of the modified ADABOC in Figure 4 by comparing the forecasted values with the original test dataset values and the ADABOC forecasted values since the proposed algorithm is a modification of it.



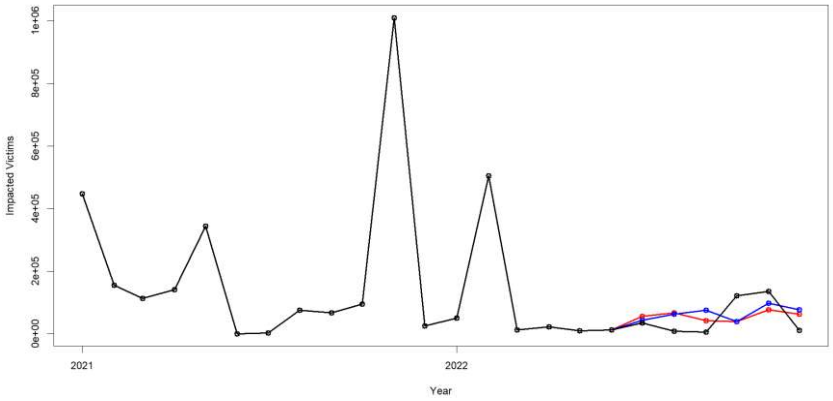
a) Education



b) Finance



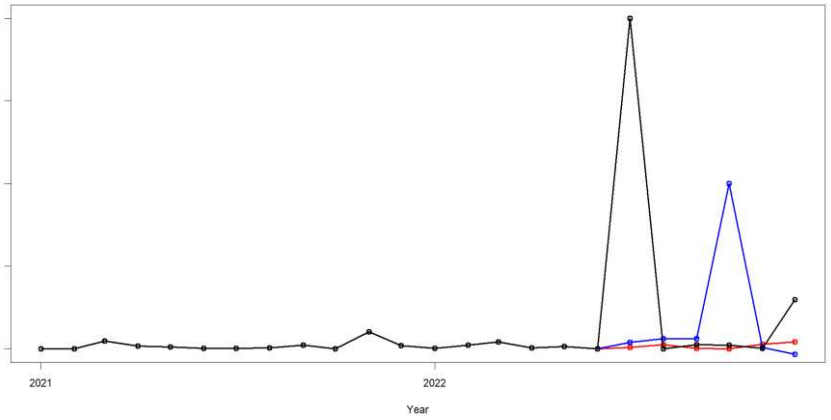
c) Health



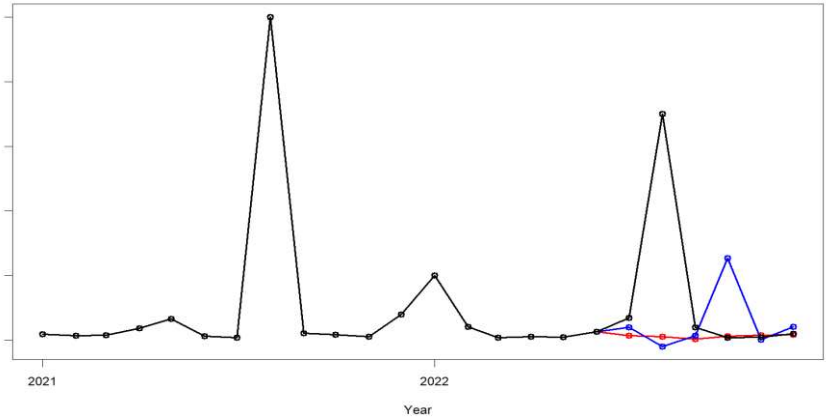
d) NGO

(Figure continues)

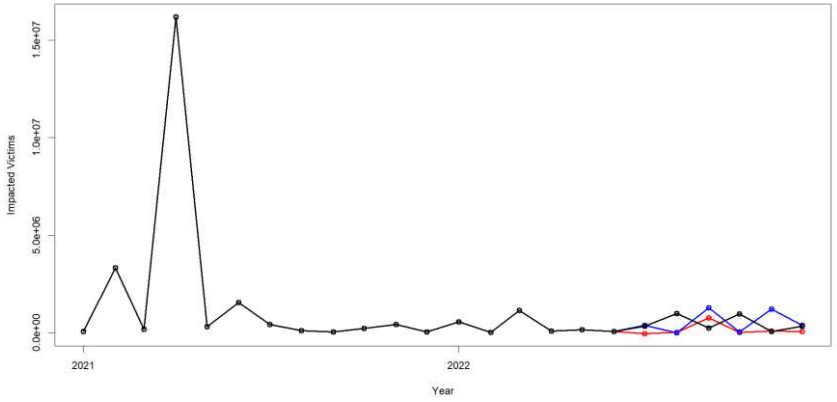
(Figure continued)



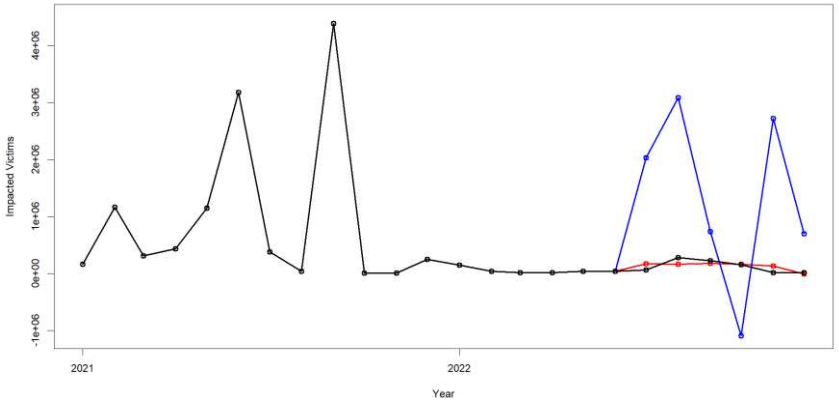
e) Hospitality



f) Manufacturing & Utilities



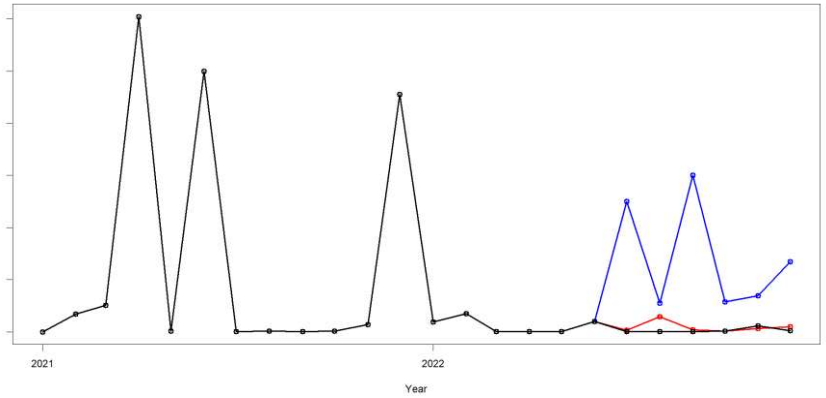
g) Professional services



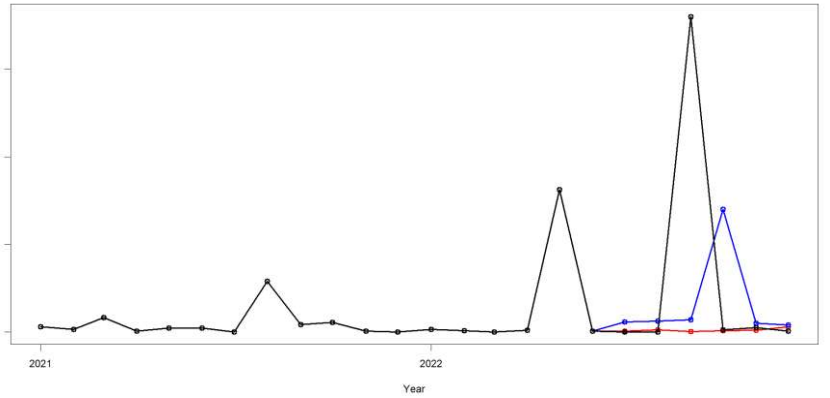
h) Retail

(Figure continues)

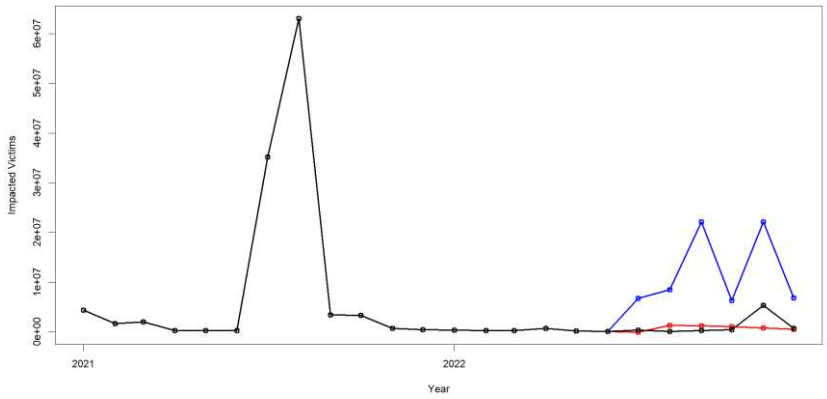
(Figure continued)



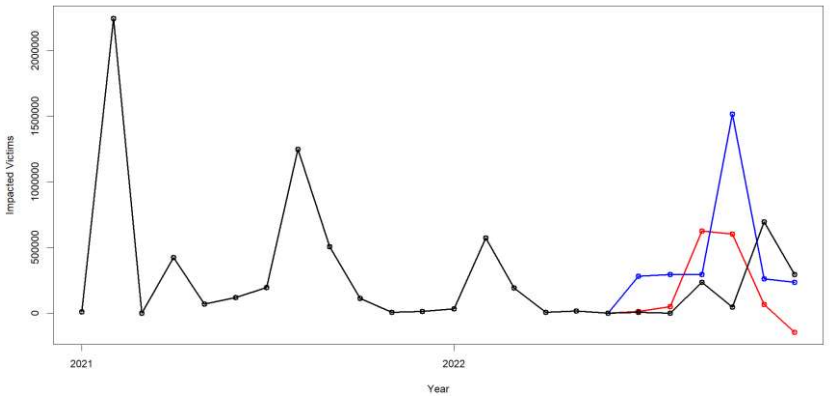
i) Technology



j) Transportation



k) Other



l) Gov't/military

Figure 4: Time series plots of all sectors from 2021 to 2022 (original — black) together with forecasts of the modified ADABOC models (red), and ADABOC models (blue)

The y-coordinates for hospitality, manufacturing & utilities, technology, and transportation from Figure 4 are omitted since the extreme values were adjusted for clear visualization of the points in the plots.

The flexibility of the modified ADABOC to select between the two averages (mean and median) during the simulation stage of the algorithm controls the effect of outliers in making forecasts while the ADABOC is heavily affected by outliers due to its conditional expectation of the simulated values. Except for healthcare and NGO sectors, we averaged the simulated values of the other sectors using the median since those sectors had their means significantly greater than their third quartiles. Thus, it is observed that the ADABOC has extreme forecasted values especially in hospitality (e), manufacturing & utilities (f), retail (h) which even has one extreme negative value, and technology (i) & *other* (k) where all six forecasted values are far above the original values. On the other hand, the modified ADABOC has close forecasts except in cases where there was an occasional spike in the number of impacted victims. Therefore, the proposed algorithm outperforms the ADABOC, the VAR, and the regression models.

CHAPTER VI: CONCLUSION AND DISCUSSION

In this study, we proposed an iterative algorithm which is a modification of the already existing ADABOC algorithm. The modifications made were mainly to average the simulated values generated through the bicopula conditional simulation conditioned on the selected variables using either the mean or the median (if there are outliers), and to sample 50% of the predictor variables during the copula selection stage to increase the probability of other variables to be selected. 50% is a tuning parameter which controls the variable selection by decorrelating the variables and the error terms to ensure other copula families and variables are selected. These changes have significantly improved the prediction performance of the ADABOC algorithm to the extent that the proposed algorithm outperforms some regression algorithms and the VAR model. The proposed algorithm is applicable to regression and other multivariate time series data.

The proposed algorithm helps estimate the number of data breach impacted victims in the various sectors. The selection of variables to explain error variables also informs sectors of other relevant sectors significant in estimating their number of impacted victims, thus, having the various sectors take cooperative measures to reduce the effect of data breach in their respective sectors.

The study has some limitations. It is limited to the ITRC identity theft data from the United States; thus, it is possible that other data from other countries may result in different estimations. Also, due to the sampling, it is important to set seed when developing the model in order to retain the variables selected, otherwise, new sets of variables may be selected in another model.

Going forward, the continuous ranked probability score will be included in measuring the prediction accuracy of the algorithm to improve the algorithm.

REFERENCES

1. BIS, (2016). *Guidance on cyber resilience for financial market infrastructures*. Bank of International Settlements (BIS).
2. Carrillo, J. A., Nieto, M., Velez, J. F., & Velez, D. (2021). A new machine learning forecasting algorithm based on bivariate copula functions. *Forecasting*, *3*(2), 355-376.
3. Chang, B., & Joe, H. (2019). Prediction based on conditional distributions of vine copulas. *Computational Statistics & Data Analysis*, *139*, 45-63.
4. Chen, L., & Guo, S. (2019). *Copulas and its application in hydrology and water resources*. Springer Singapore.
5. DeMatteis, R. (2001). *Fitting copulas to data. Institute of Mathematics of the University of Zurich* (Doctoral dissertation, Diploma Thesis).
6. Federal Bureau of Investigation, (2022). *Internet crime report*. Retrieved from https://www.gasa.org/post/fbi-internet-crime-report-2022
7. Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, *50*(4), 361-365.
8. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: Springer.
9. Macaulay, T., & Centre for International Governance. (2019). The Danger of Critical Infrastructure Interdependency. In *Governing Cyberspace during a Crisis in Trust: An essay series on the economic potential — and vulnerability — of transformative technologies and cyber security* (pp. 69–73). Centre for International Governance Innovation. http://www.jstor.org/stable/resrep26129.16
10. Nelsen, R. B. (2006). *An introduction to copulas*. Springer, New York. *MR2197664*.
11. NVIDIA (2023). *XGBoost*. NVIDIA Corporation.

12. Peng, C., Xu, M., Xu, S., & Hu, T. (2018). Modeling multivariate cybersecurity risks. *Journal of Applied Statistics*, *45*(15), 2718-2740.

13. Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP* (Vol. 8, No. 3, pp. 229-231).

14. Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., Nagler, T., & Erhardt, T. (2015). *Package 'VineCopula'*. R package version 1.6-1.

15. Tsay, R. S. (2005). *Analysis of financial time series*. John Wiley & Sons.

16. Venter, G. (2001). *Tails of Copulas*. ASTIN: Washington, DC, USA, pp. 68–113.

17. Xu, M., Hua, L., & Xu, S. (2017). A vine copula model for predicting the effectiveness of cyber defense early-warning. *Technometrics*, *59*(4), 508-520.

18. Zhang Wu, M., Luo, J., Fang, X., Xu, M., & Zhao, P. (2021). Modeling multivariate cyber risks: deep learning dating extreme value theory. *Journal of Applied Statistics*, 1-21.

APPENDIX: BIVARIATE COPULA FAMILIES IN THE R-VINE COPULA PACKAGE USED IN THE MODIFIED ADABOC AND ADABOC

independence copula (∏)
Gaussian copula
Student t copula (t-copula)
Clayton copula
Gumbel copula
Frank copula
Joe copula
BB1 copula
BB6 copula
BB7 copula
BB8 copula
rotated Clayton copula (180 degrees; survival Clayton")
rotated Gumbel copula (180 degrees; survival Gumbel")
rotated Joe copula (180 degrees; survival Joe")
rotated BB1 copula (180 degrees; survival BB1")
rotated BB6 copula (180 degrees; survival BB6")
rotated BB7 copula (180 degrees; survival BB7")
rotated BB8 copula (180 degrees; "survival BB8")
rotated Clayton copula (90 degrees)
rotated Gumbel copula (90 degrees)
rotated Joe copula (90 degrees)
rotated BB1 copula (90 degrees)
rotated BB6 copula (90 degrees)
rotated BB7 copula (90 degrees)
rotated BB8 copula (90 degrees)
rotated Clayton copula (270 degrees)
rotated Gumbel copula (270 degrees)
rotated Joe copula (270 degrees)
rotated BB1 copula (270 degrees)
rotated BB6 copula (270 degrees)
rotated BB7 copula (270 degrees)
rotated BB8 copula (270 degrees)
Tawn type 1 copula
rotated Tawn type 1 copula (180 degrees)
rotated Tawn type 1 copula (90 degrees)
rotated Tawn type 1 copula (270 degrees)
Tawn type 2 copula
rotated Tawn type 2 copula (180 degrees)
rotated Tawn type 2 copula (90 degrees)
rotated Tawn type 2 copula (270 degrees)