ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 12, Nomor 3, Tahun 2023, Halaman 392 - 402
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PENGGUNAAN *MIXTURE MODEL* KERNEL-*GENERALIZED PARETO DISTRIBUTION* DAN D-VINE COPULA DALAM MENGANALISIS UKURAN PELANGGARAN DATA

Fathiyyah Yolianda Dzikra[1*], Yuciana Wilandari[2], Arief Rachman Hakim[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail**: fathiyyahyd@gmail.com

DOI: 10.14710/j.gauss.12.3.392-402

Article Info:
Received: 2022-12-12
Accepted: 2024-02-13
Available Online: 2024-02-26

Keywords:
Breach Sizes; *Generalized Pareto Distribution;* Kernel; D-Vine Copula

Abstract: The research conducted on the 2015-2021 Data Breach Report in the U.S. Department of Health and Human Services is a study related to the estimation and modeling of the breach sizes each type of entity using the Kernel-Generalized Pareto Distribution Mixture Model method, as well as the estimation of the dependence of breach sizes between years with the D-Vine Copula. The D-Vine Copula can accommodate the complex dependencies demonstrated by data breach reports across all enterprise categories. Before researching with D-Vine Copula, we will first model and estimate breach size parameters for each type of entity using the Mixture Model Kernel-Generalized Pareto Distribution (GPD). The Mixture Model can accommodate large data breach sizes via GPD and also allows the use of non-parametric kernel distributions to model smaller data breach sizes. The data resulting from the logarithmic transformation of entity data in the Business Associate and Healthcare Provider types has a right short-tail with Weibull distribution, while the Health Plan category has a right heavy-tail with Frechet distribution. The three types of entity were estimated using the maximum likelihood Cross-Validation method. Dependency estimation with D-Vine Copula shows that the breach sizes between years measure has a positive dependency.

1. PENDAHULUAN

Data dari sebuah perusahaan didapatkan dari berbagai sumber, diantaranya data yang diperoleh dengan persetujuan klien, data yang didapatkan dari informasi publik, data yang berasal dari kerjasama antar perusahaan, dan lain-lain. Data-data yang sudah didapat tentu saja dijaga oleh perusahaan tersebut dengan ketat untuk menghindari terjadinya pelanggaran data. Namun sebaik-baik dan seketat-ketatnya perusahaan menjaga datanya, masih terdapat banyak hal yang menyebabkan pelanggaran data terjadi.

Organisasi *Identity Theft Resource Center* (2021) mengungkapkan bahwa, bulan Mei tahun 2021 merupakan bulan dengan kasus pelanggaran data terbanyak dalam 5 tahun terakhir di Amerika Serikat. Fang *et al* (2021) meneliti menggunakan sebuah kerangka kerja untuk memodelkan dan memprediksi *multivariate time series* terhadap pelanggaran data yang jarang terjadi. Hasil penelitian tersebut memiliki kesimpulan bahwa metode D-Vine Copula bisa mengakomodasi dependensi kompleks yang ditunjukkan oleh laporan pelanggaran data di semua kategori perusahaan.

Penelitian ini meneliti pelanggaran data di bidang kesehatan dan pelayanan umum Amerika Serikat yang dilaporkan dari Januari 2015 hingga Oktober 2021 dengan tujuan mengestimasi parameter ukuran pelanggaran data tiap kategori perusahaan dengan metode *Mixture Model* Kernel-GPD, serta memodelkan ukuran pelanggaran data tiap kategori

perusahaan berdasarkan parameter yang telah didapat, lalu mengestimasi dependensi ukuran pelanggaran data antar tahun dengan metode D-Vine Copula.

Mixture Model Kernel-GPD merupakan model gabungan antara distribusi kernel dan distribusi *Generalized Pareto* (GPD). Model ini dikembangkan dengan tujuan agar proses analisis nilai-nilai ekstrim menjadi lebih fleksibel. Menurut Sun *et al* (2020), *mixture model* Kernel-GPD sangat fleksibel untuk digunakan dalam memodelkan ukuran pelanggaran data, dimana *mixture model* dapat mengakomodasi kondisi *heavy tail* yang terjadi pada ukuran pelanggaran data yang besar melalui GPD dan juga memungkinkan penggunaan distribusi kernel untuk memodelkan ukuran pelanggaran data yang lebih kecil.

Metode copula menggabungkan beberapa variabel dengan fungsi distribusi berbeda ke dalam bentuk distribusi bersama. Copula sepenuhnya menangkap struktur dependensi di antara variabel acak dan bersifat fleksibel dimana tidak memerlukan asumsi normalitas. Metode copula juga dapat melihat struktur dependensi antar variabel dengan distribusi yang berbeda. Pemodelan menggunakan copula vine menciptakan copula multivariat yang menggunakan copula bivariat sebagai landasannya. Salah satu kelas dari copula vine adalah copula d-vine yang mempunyai struktur berurut. Estimasi parameter dari copula d-vine dapat menunjukkan bagaimana dependensi ukuran pelanggaran data antar tahun pada data yang akan diteliti.

2. TINJAUAN PUSTAKA

Transformasi data dilakukan apabila data yang akan diteliti tidak berdistribusi normal, memiliki banyak ragam, ataupun keduanya. Transformasi jenis logaritma sering diterapkan pada data - data yang berupa ukuran (*size*), memiliki nilai yang besar, memiliki ragam yang banyak, tidak berdistribusi normal, atau memiliki kondisi keruncingan yang parah. Transformasi logaritma mereduksi kemiringan data dan sering cocok untuk digunakan pada data yang akan dianalisis.

Menurut Friederichs (2007), *Extreme Value Theory* (EVT) berfokus pada nilai ekstrim serta kejadian yang jarang terjadi. Metode *Peak Over Threshold* (POT) merupakan salah satu metode pendekatan dalam EVT yang bisa digunakan untuk menganalisis nilai ekstrim. Metode POT adalah salah satu metode untuk menganalisis nilai ekstrim dengan melihat nilai ekstrim dari suatu sampel yang melampaui ambang batas atau *threshold* tertentu.

Threshold merupakan batas ambang dalam menentukan nilai ekstrim, dimana nilai ekstrim adalah nilai-nilai yang lebih besar dari *threshold* yang ditentukan. Secara teoritis, metode yang umum dan juga lebih mudah digunakan dalam menentukan *threshold* adalah metode persentase, yaitu dengan melihat nilai persentil dari data. Persentil ke 90 digunakan sebagai *threshold*, sehingga 10% data di atasnya merupakan nilai ekstrim.

Berdasarkan Kang dan Song (2017), teorema Pickands-Balkema-de Haan menyatakan bahwa distribusi untuk nilai ekstrim, secara umum diasumsikan berdistribusi *Generalized Pareto* (GPD). GPD memiliki bentuk fungsi distribusi sebagai berikut (Hu, 2013).

$$G(y|\mu,\sigma_u,\xi)=\begin{cases}1-\left[1+\xi\left(\frac{y-\mu}{\sigma_u}\right)\right]^{-1/\xi} & \xi\neq 0\\ 1-exp\left[-\left(\frac{y-\mu}{\sigma_u}\right)\right] & \xi=0.\end{cases} \quad (1)$$

y adalah nilai ekstrim, μ merupakan parameter lokasi yang sering dianggap bernilai 0, ξ merupakan parameter bentuk, dan σ_u merupakan parameter skala sedangkan u merupakan

notasi *threshold*. Parameter-parameter GPD diestimasi menggunakan metode maksimum likelihood beserta metode Newton-Raphson sebagai bantuan.

Data non-ekstrim dianalisis dengan distribusi kernel yang memiliki fungsi densitas sebagai berikut (Herawati, 2017).

$$r(x; x_h, \lambda) = \frac{1}{l\lambda}\sum_{h=1}^{l} K\left(\frac{x - x_h}{\lambda}\right), \tag{2}$$

dimana x adalah variabel bebas, x_h adalah sampel ke-h dimana $h = 1, \dots, l$, l adalah banyaknya data non-ekstrim, K adalah fungsi kernel, dan λ adalah nilai *bandwidth*. *Bandwidth* adalah parameter *smoothing* bernilai positif yang digunakan untuk mengatur derajat kehalusan. Nilai *bandwidth* diestimasi menggunakan metode maksimum likelihood *cross-validation* (MLCV). Nilai *bandwidth* yang baik didefinisikan memiliki nilai yang mendekati batas maksimum, yaitu,

$$\lambda_{mlcv} = \underset{\lambda>0}{\operatorname{argmax}}\, \mathrm{MLCV}(\lambda), \tag{3}$$

dimana MLCV adalah nilai maksimum likelihood *cross validation* yang digunakan untuk memilih jenis kernel yang akan digunakan. Pemilihan dilihat berdasarkan nilai MLCV yang paling kecil. MLCV memiliki rumus persamaan sebagai berikut.

$$\mathrm{MLCV}(\lambda) = \frac{1}{l}\sum_{g=1}^{l} \log\left(\sum_{\substack{h=1 \\ h\neq g}}^{l} K_\lambda(x_g - x_h)\right) - \log\left(\frac{l-1}{\lambda}\right), \tag{4}$$

MacDonald et al (2011) menciptakan *Mixture Model* Kernel-GPD. Model ini dikembangkan dengan tujuan agar proses analisis nilai-nilai ekstrim menjadi lebih fleksibel. *Mixture model* Kernel-GPD merupakan model gabungan antara GPD dan distribusi kernel. Fungsi distribusi model campuran Kernel-GPD adalah sebagai berikut,

$$M(x|x_h, \lambda, \mu, \sigma_u, \xi, \phi_u) = \begin{cases} (1-\phi_u)\frac{R(x|x_h,\lambda)}{R(u|x_h,\lambda)}, & x \le u, \\ (1-\phi_u) + \phi_u G(x|\mu, \sigma_u, \xi), & x > u, \end{cases} \tag{5}$$

dimana $R(.\,|x_h, \lambda)$ adalah fungsi distribusi kernel dan $G(x|\mu, \sigma_u, \xi)$ menggambarkan fungsi GPD. Sedangkan ϕ_u adalah proporsi untuk data yang lebih besar dari *threshold*, bernilai $0 < \phi_u < 1$, dan berasal dari perhitungan $\phi_v = 1 - R(u|x_h, \lambda)$.

Copula berasal dari bahasa latin yang berdasarkan Kamus Bahasa Latin Cassell's memiliki arti dalam bahasa inggris *link*, *tie*, *bond*. Czado dan Nagler (2022) menyampaikan bahwa pendekatan dengan copula memungkinkan model marginal dibangun untuk masing-masing variabel secara terpisah dan variabel-variabel tersebut dapat dihubungkan dengan struktur dependensi yang digolongkan oleh copula. Kata copula pertama kali digunakan oleh Sklar pada tahun 1959, dalam teorinya yang menjadi teorema sentral dari teori copula, dan dikenal sebagai Teorema Sklar.

Misalkan F adalah fungsi distribusi multivariat d-dimensi dengan distribusi marginal $F_1, \dots, F_d$, maka terdapat sebuah copula C untuk $\boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$ sebagai berikut,

$$F(\boldsymbol{x}) = C\big(F_1(x_1), \dots, F_d(x_d)\big). \tag{6}$$

d adalah banyaknya dimensi data. Untuk setiap $d \geq 2$, copula d-dimensi adalah fungsi distribusi d-variat dengan range $[0,1]^d$ yang distribusi marginal nya berdistribusi uniform $[0,1]$. Pada dasarnya copula merupakan cara untuk mentransformasi variabel acak

$(X_1, \dots, X_d)$ menjadi variabel acak lain $(U_1, \dots, U_d) = \big(F_1(X_1), \dots, F_d(X_d)\big)$ yang distribusi marginal nya berdistribusi uniform [0,1] dan menjaga dependensi antar komponen di dalamnya (Durante dan Sempi, 2010).

$F(\boldsymbol{x})$ bersifat kontinu apabila copula C dan distribusi marginal $F_1, \dots, F_d$ juga kontinu. Dalam hal itu, maka fungsi densitas $f(\boldsymbol{x})$ dari $F(\boldsymbol{x})$ memenuhi,

$$f(\boldsymbol{x}) = c\big(F_1(x_1), \dots, F_d(x_d)\big) \prod_{t=1}^{d} f_t(x_t), \tag{7}$$

untuk $\boldsymbol{x} \in \prod_{t=1}^{d} \operatorname{ran} X_t$, dimana $\operatorname{ran} X_t = \{x \in \mathbb{R} : f_t(x) > 0\}$ untuk setiap $t \in \{1, 2, \dots, d\}$. f_t merupakan densitas dari F_t dan c adalah densitas dari copula C.

Copula Gaussian atau copula Normal merupakan copula dari $\boldsymbol{X} = (X_1, \dots, X_d)$ yang berdistribusi normal standar d-variat dengan vektor mean 0 dan matriks korelasi P yang memiliki bentuk fungsi copula sebagai berikut,

$$C_P^{Gauss}(u_1, \dots, u_d) = \Phi_P^d\big(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\big), \tag{8}$$

dimana $(u_1, \dots, u_d) \in [0,1]^d$, Φ_P^d adalah fungsi distribusi kumulatif gabungan dari distribusi normal standar d-variat dengan matriks korelasi P, dan $\Phi^{-1}(.)$ merupakan invers dari distribusi normal standar.

Copula Frank merupakan salah satu variasi fungsi pembangkit dari copula Archimedean. Berdasarkan Nelsen (2006), berikut fungsi distribusi dari copula Frank,

$$C_\delta^{Frank}(u_1, \dots, u_d) = -\frac{1}{\delta} \ln\left(1 + \frac{\prod_{t=1}^{d}\left(e^{-\delta u_t} - 1\right)}{\left(e^{-\delta} - 1\right)^{d-1}}\right). \tag{9}$$

δ adalah parameter copula frank.

Pemodelan menggunakan copula vine menciptakan copula multivariat yang menggunakan copula bivariat sebagai landasannya. Czado (2010) mengatakan bahwa, titik awal dari membangun distribusi multivariat adalah dekomposisi pengulangan dari densitas multivariat menjadi densitas bersyarat. Aas *et al* (2009) melambangkan fungsi distribusi bersyarat dalam kondisi x dan z bersifat uniform sebagai fungsi $h(x, z, \Theta)$ dengan persamaan sebagai berikut,

$$h(x, z, \Theta) = F(x|z) = \frac{\partial C_{xz}(F(x), F(z))}{\partial\, F(z)}, \tag{10}$$

Θ menunjukkan kumpulan parameter dari fungsi distribusi gabungan copula dari x dan z.

Fungsi densitas $f(x_1, \dots, x_d)$ yang disebut Bedford dan Cooke (2002) sebagai dekomposisi distribusi D-vine memiliki bentuk sebagai berikut

$$f(x_1, \dots, x_d) = \left[\prod_{t=1}^{d-1} \prod_{s=1}^{d-t} c_{s,(s+t)|(s+1),\dots,(s+t-1)}\right] \cdot \left[\prod_{j=1}^{d} f_j(x_j)\right], \tag{11}$$

$$^{*}c_{s,(s+t)|s+1,\dots,(s+t)-1} \coloneqq c_{s,(s+t)|s+1,\dots,(s+t)-1}\Big(F\big(x_s|x_{s+1}, \dots, x_{(s+t)-1}\big), F\big(x_{(s+t)}|x_{s+1}, \dots, x_{(s+t)-1}\big)\Big).$$

Copula vine berdimensi-d mempunyai $d - 1$ pohon. Ciri khas dari pohon D-vine adalah, masing-masing simpul hanya terhubung maksimal pada 2 simpul lainnya. 2 simpul yang terhubung menghasilkan 1 tepi. Menurut Fang *et al* (2021), setiap tepi terkait dengan densitas pasangan-copula yang digunakan untuk memodelkan dependensi antara 2 variabel, lalu label tepi menunjukkan parameter dependensi dari densitas pasangan-copula terkait.

Parameter copula multivariat diestimasi menggunakan metode maksimum likelihood dengan fungsi log likelihood D-vine adalah sebagai berikut,

$$\ell(x_{it}) = \sum_{i=1}^{n}\sum_{t=1}^{d-1}\sum_{s=1}^{d-t} \log\left(c_{i;\, s,(s+t)|(s+1),\ldots,(s+t-1)}\right) + \sum_{i=1}^{n}\sum_{j=1}^{d} \log\left(f_{i,j}(x_{ij})\right). \quad (12)$$

Sedangkan parameter untuk copula bivariat memiliki fungsi log likelihood sebagai berikut,

$$\ell(\boldsymbol{u}_i|\theta) = \log c_{st}(u_{is}, u_{it}, \theta_{st}), \quad (13)$$

dimana $s, t = 1, \ldots d\ s \neq t$ dan $i = 1, .. n.\ u_{is}/u_{it}$ merupakan data asli atau data semu yang berdistribusi uniform. Fungsi log likelihood densitas copula bivariat untuk copula gaussian adalah sebagai berikut,

$$\ell_{\rho}^{Gauss}(u_{i1}, u_{i2}) = -\frac{1}{2}\log(1-\rho^2) + \frac{2\rho x_1 x_2 - {x_1}^2 - {x_2}^2}{2(1-\rho^2)} + \frac{{x_1}^2 + {x_2}^2}{2}, \quad (14)$$

dimana $x_1 = \Phi^{-1}(u_{i1})$ dan $x_2 = \Phi^{-1}(u_{i2})$ dan ρ adalah parameter copula gaussian dimana $\rho \in [-1,1]$. Sedangkan Fungsi log likelihood densitas copula bivariat untuk copula frank dengan δ sebagai parameter copula frank adalah sebagai berikut.

$$\ell_{\delta}^{Frank}(u_{i1}, u_{i2}) = 2\log\left[\left(e^{-\delta u_{i1}} - 1\right)\left(e^{-\delta u_{i2}} - 1\right) + \left(e^{-\delta} - 1\right)\right]. \quad (15)$$

Metode *Akaike Information Criteria* (AIC) digunakan untuk menentukan model copula terbaik yang pengestimasian parameternya menggunakan metode maksimum likelihood dengan rumus sebagai berikut,

$$AIC = -2(\log likelihood) + 2\theta, \quad (16)$$

dimana θ adalah jumlah parameter model copula. Jika nilai AIC semakin kecil, maka semakin baik pula modelnya.

Menurut bagian Administrasi Anak dan Keluarga Departemen Kesehatan dan Pelayanan Umum Amerika Serikat (2015), pelanggaran data atau *data breach* adalah pelanggaran keamanan dimana data rahasia, sensitif atau data yang dilindungi, disalin, dikirim, dilihat, diambil, atau digunakan oleh pihak yang tidak memiliki kewenangan atas data tersebut. Departemen Kesehatan dan Pelayanan Umum Amerika Serikat (HHS) mengungkapkan bahwa sejak tahun 2009 hingga tahun 2014, laporan kasus pelanggaran data terus meningkat. Walaupun sempat turun pada tahun 2015, laporan kasus pelanggaran data kembali meningkat pada tahun 2016 hingga tahun 2019. HHS membagi jenis pelanggaran menjadi 5 tipe, yaitu peretasan/insiden IT, pencurian, kehilangan, kecerobohan dalam pembuangan data tak terpakai, dan pengaksesan yang tidak sah.

Pelanggaran data memberikan dampak bagi pihak klien maupun perusahaan. Bagi pihak klien tentu saja akan terjadi penyalahgunaan data, dimana data klien bisa saja tersebar dan disalahgunakan oleh pihak-pihak tertentu. Bagi perusahaan, Organisasi *Cypress Data Defense* mengatakan terdapat 2 jenis dampak finansial bagi perusahaan yaitu dampak langsung dan tidak langsung. Dampak langsung bagi perusahaan adalah pengokohan sistem keamanan, membayar denda, dan menyelesaikan tuntutan hukum, sedangkan dampak tidak langsung adalah kehilangan klien, penurunan pendapatan karena rusaknya reputasi perusahaan, dan lain-lain.

3. METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah data sekunder berupa Laporan Kasus Pelanggaran Data di Bidang Kesehatan dan Pelayanan Umum Amerika Serikat Periode Januari 2015 hingga Oktober 2021 yang diperoleh dari website Departemen Kesehatan dan

Pelayanan Umum Amerika Serikat bagian arsip laporan pelanggaran (*Breach Report*) pada 11 Januari 2022.

Variabel yang diteliti adalah ukuran pelanggaran data atau *Breach Sizes* dengan mengelompokkan ukuran pelanggaran data dari 2 sisi. Yang pertama dari sisi kategori perusahaan, dimana ada 3 kategori yaitu *Business Associate* (BA), *Health Plan* (HP) dan *Healthcare Provider* (HCP). Sedangkan sisi lainnya adalah dari sisi tahun kasus pelanggaran data dilaporkan, sehingga terdapat 7 periode tahun yaitu dari 2015 hingga 2021.

Berikut merupakan tahapan analisis data pada penelitian ini:

1. Mengambil data dari website *Breach Report* milik Departemen Kesehatan dan Pelayanan Umum Amerika Serikat
2. Mentransformasi ukuran pelanggaran data menggunakan fungsi logaritma
3. Mencari *threshold* dan memisahkan antara data ekstrim dan non-ekstrim
4. Mengestimasi parameter data ekstrim dengan GPD
5. Mengestimasi parameter *bandwidth* data non-ekstrim dengan distribusi kernel
6. Pemodelan campuran berdasarkan parameter GPD dan distribusi kernel yang sudah didapatkan
7. Hasil pemodelan dibawa ke bentuk CDF dan ditransformasi ke bentuk distribusi uniform.
8. Data hasil transformasi uniform dibuat ke dalam bentuk matriks $d \times n$, d adalah dimensi data periode tahun dan n adalah banyaknya perusahaan.
9. Menentukan matriks struktur copula d-vine dan membangun konstruksi copula
10. Mengestimasi parameter copula yang dimulai dengan mengestimasi parameter-parameter copula bivariat di pohon pertama dengan menggunakan data asli yang berdistribusi uniform.
11. Menghitung data semu untuk pohon kedua dengan fungsi distribusi bersyarat yang diperoleh dari perhitungan Persamaan (10) yang sesuai dan parameter copula pohon pertama.
12. Mengulangi langkah 10 dan 11 hingga pohon terakhir dalam konstruksi copula untuk mendapatkan semua parameter copula.
13. Memilih copula terbaik.

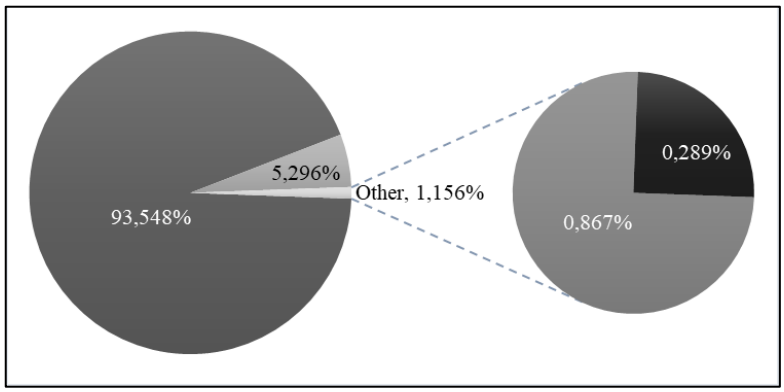
4. HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini adalah laporan pelanggaran data di bidang kesehatan dan pelayanan umum Amerika Serikat yang diambil pada 11 Januari 2022, yang memuat kasus pelanggaran data yang dilaporkan dari bulan Januari 2015 hingga bulan Oktober 2021. Data tersebut terdiri dari 2243 kasus pelanggaran data yang terbagi dalam 3 kategori perusahaan yaitu *Business Associate* (BA), *Health Plan* (HP), dan *Healthcare Provider* (HCP).

Tabel 1. Statistik Deskriptif Data Asli

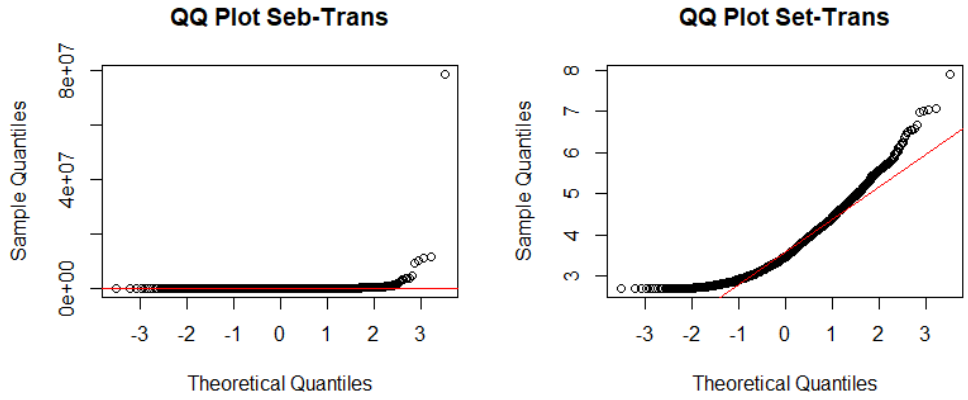
Type	Enterprise	Records	Year
Business Associate: 188	Length:2243	Min. : 500	2015:261
Health Plan: 305	Class :character	1st Qu.: 1078	2016:317
Healthcare Provider:1750	Mode :character	Median : 2933	2017:347
		Mean : 93297	2018:349
		3rd Qu.: 12654	2019:470
		Max. :78800000	2020:413
		Sd : 1735430	2021: 86

Berdasarkan Tabel 1, kasus pelanggaran data paling banyak terjadi pada perusahaan dengan kategori HCP yaitu 1750 kasus. Dari kolom *Records* yang merupakan ukuran pelanggaran data (*breach sizes*), dapat disimpulkan bahwa terdapat jangkauan serta varian data yang sangat besar. Jumlah kasus pelanggaran data antar tahun terpantau terus naik hingga mencapai puncaknya tahun 2019.



Gambar 1. Kasus Pelanggaran Data dari Sisi Perusahaan

Sebanyak 2243 kasus pelanggaran data menimpa 2077 perusahaan, dimana 0,289% diantaranya mengalami lebih dari 3 kali pelanggaran data, 0,818% diantaranya mengalami 3 kali pelanggaran data, 5,392% mengalami 2 kali pelanggaran data, lalu sisanya mengalami 1 kali pelanggaran data.



Gambar 2. Plot Data Sebelum dan Setelah Transformasi Logaritma

Berdasarkan Gambar 2 dapat terlihat plot-plot data yang awalnya landai berubah ke bentuk plot *right skewed*. Kondisi tersebut didukung oleh Tabel 2 yang menampilkan nilai *skewness* data keseluruhan maupun per kategori perusahaan bernilai positif, sehingga data hasil transformasi dapat dikatakan miring ke kanan atau *right skewed*. Dari Tabel 2 juga terlihat bahwa nilai standar deviasi sudah sangat lebih kecil dari nilai standar deviasi sebelum transformasi, hal ini menunjukkan bahwa data tersebar lebih merata dari sebelumnya.

Tabel 2. Statistik Deskriptif Data Hasil Transformasi

	Mean	SD	Skewness	Kurtosis	25%	50%	75%	n
BA	3,8431	0,9104	1,0723	1,0433	3,1347	3.6357	4.3186	188
HP	3,6280	0,8219	1,7036	4,1255	3,0000	3.4234	3.9687	305
HCP	3,6304	0,7387	0,9468	0,5256	3,0291	3.4646	4.0923	1750
Records	3,6479	0.7681	1.1153	1.3659	3.0328	3.4673	4.1022	2243

Threshold (u) merupakan bagian dari parameter GPD. Setelah menentukan besaran *threshold*, metode maksimum likelihood diterapkan pada data ekstrim untuk mencari parameter skala (σ_u) dan bentuk (ξ).

Tabel 3. Estimasi Parameter GPD

	Parameter	σ_u	ξ	u (k)
BA	Estimasi	1,3815	-0,6239	4,9706 (19)
	SE	0,4173	0,2431	-
HP	Estimasi	0,5520	0,2742	4,6973 (31)
	SE	0,2236	0,3664	-
HCP	Estimasi	0,5717	-0,1760	4,6990 (173)
	SE	0,0527	0,0538	-

Notasi k dalam Tabel 3 merupakan banyaknya data ekstrim. Dari Tabel 3 diketahui perusahaan kategori BA dan HCP dengan nilai parameter bentuk yang negatif memiliki ekor kanan yang kurus atau *short-tailed* dan berdistribusi weibull. Sedangkan untuk perusahaan kategori HP memiliki parameter bentuk bernilai positif yang menunjukkan bahwa ekor kanan data bersifat gemuk atau *heavy-tailed* dan berdistribusi frechet atau pareto.

Tabel 4. Nilai MLCV tiap Jenis Kernel

	Nilai MLCV BA	Nilai MLCV HP	Nilai MLCV HCP
Gaussian	-0.8507	-0.6459	-0.6725
Epanechnikov	-0.8471	-0.6309	-0.6474
Uniform	-0.8424	-0.6315	-0.6583
Triagular	-0.8333	-0.6284	-0.6425
Biweight	-0.8362	-0.6290	-0.6426

Berdasarkan Tabel 4, maka jenis kernel yang akan digunakan untuk perusahaan katerogi BA, HP, maupun HCP adalah kernel Gaussian yang memiliki nilai MLCV terkecil diantara jenis kernel yang lain.

Dalam Tabel 5, l merupakan banyaknya data non-ekstrim dan λ merupakan notasi untuk *bandwidth*. Meskipun memiliki perbedaan dalam besaran nilai MLCV dan banyaknya data non-ekstrim, ketiga kategori perusahaan memiliki nilai *bandwidth* yang sama ketika diestimasi menggunakan metode maksimum likelihood *Cross-Validation* dengan jenis kernel Gaussian.

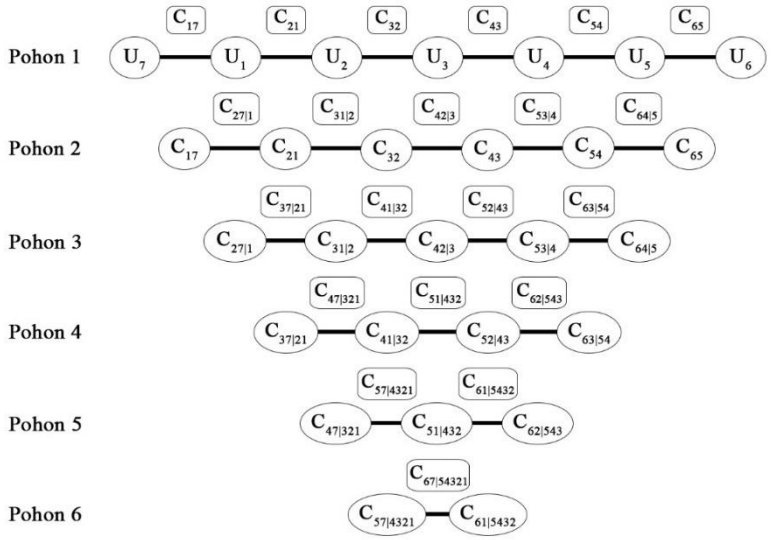
Tabel 5. Estimasi Parameter Kernel

	λ	MLCV	l
BA	0,1058	-0,8507	169
HP	0,1058	-0,6459	274
HCP	0,1058	-0,6725	1574

Matriks struktur copula d-vine yang digunakan adalah sebagai berikut.

$$\begin{bmatrix} 7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 6 & 1 & 0 & 0 & 0 & 0 & 0 \\ 5 & 6 & 2 & 0 & 0 & 0 & 0 \\ 4 & 5 & 6 & 3 & 0 & 0 & 0 \\ 3 & 4 & 5 & 6 & 4 & 0 & 0 \\ 2 & 3 & 4 & 5 & 6 & 5 & 0 \\ 1 & 2 & 3 & 4 & 5 & 6 & 6 \end{bmatrix}$$

Sehingga terbentuklah konstruksi copula d-vine yang ditunjukkan oleh Gambar 3.



Gambar 3. Konstruksi Copula D-Vine yang Digunakan

Tabel 6. Estimasi Parameter Copula

Pohon	Pasangan	Gaussian		Frank	
		Parameter (ρ)	Tau (τ)	Parameter (δ)	Tau(τ)
1	1,7	0,65	0,45	11,88	0,71
	2,1	0,55	0,37	7,21	0,57
	3,2	0,52	0,35	6,13	0,52
	4,3	0,50	0,34	5,76	0,50
	5,4	0,47	0,31	4,95	0,45
	6,5	0,45	0,30	4,58	0,43
2	2,7;1	0,36	0,23	5,87	0,51
	3,1;2	0,29	0,19	3,79	0,37
	4,2;3	0,28	0,18	3,27	0,33
	5,3;4	0,24	0,16	2,61	0,27
	6,4;5	0,29	0,19	2,80	0,29
3	3,7;2,1	0,29	0,19	4,98	0,45
	4,1;3,2	0,23	0,14	3,27	0,33
	5,2;4,3	0,18	0,11	2,14	0,23
	6,3;5,4	0,21	0,13	2,28	0,24
4	4,7;3,2,1	0,26	0,17	4,71	0,44
	5,1;4,3,2	0,16	0,10	2,35	0,25
	6,2;5,4,3	0,17	0,11	2,20	0,23
5	5,7;4,3,2,1	0,22	0,14	3,50	0,35
	6,1;5,4,3,2	0,16	0,10	2,51	0,26
6	6,7;5,4,3,2,1	0,23	0,15	3,92	0,38
Log Likelihood			142857,9		22415,78
AIC			-285673,7		-44789,57

Berdasarkan nilai AIC pada Tabel 6, dapat disimpulkan, copula terbaik adalah copula gaussian, karena copula tersebut memiliki nilai AIC yang paling kecil yaitu -285673,7. Tabel 6 juga menampilkan nilai parameter yang positif di semua pasangan copula. Sehingga dapat disimpulkan bahwa ukuran pelanggaran data antar tahun pada laporan pelanggaran data di

bidang kesehatan dan pelayanan umum Amerika Serikat tahun 2015-2021 memiliki dependensi yang positif, hal ini juga didukung oleh nilai tau tiap pasangan copula yang positif. Dependensi positif pada ukuran pelanggaran data dapat diartikan sebagai, jika di masa lampau sebuah perusahaan tidak terkena pelanggaran data, maka di masa depan pun pelanggaran data diantisipasi tidak akan terjadi, dan apabila sebuah perusahaan terkena pelanggaran data di masa lampau, kemungkinan di masa depan pelanggaran data akan terjadi pada perusahaan tersebut.

5. KESIMPULAN

Metode *mixture model* kernel-GPD digunakan untuk mengestimasi paramater ukuran pelanggaran data per kategori perusahaan. Data asli terlebih dahulu ditransformasi menggunakan fungsi logaritma. GPD diterapkan pada data ekstrim dan distribusi kernel diterapkan pada data non ekstrim. Hasil estimasi parameter GPD pada data non ekstrim memiliki hasil bahwa data perusahaan kategori *Business Associate* dan *Healthcare Provider* memiliki ekor kanan yang kurus dan berdistribusi weibull. Sedangkan untuk perusahaan kategori *Health Plan* menunjukkan bahwa ekor kanan data bersifat gemuk dan berdistribusi frechet atau pareto. Untuk distribusi kernel pada data non-ekstrim, dengan jenis kernel yang terpilih yaitu kernel gaussian, dperoleh hasil bahwa ketiga kategori perusahaan memiliki nilai *bandwidth* yang sama. Estimasi dependensi pada ukuran pelanggaran data dengan metode D-Vine Copula memperoleh hasil copula gaussian sebagai copula terbaik dan dapat disimpulkan bahwa ukuran pelanggaran data antar tahun memiliki dependensi yang positif. Dependensi positif pada ukuran pelanggaran data dapat diartikan sebagai, jika di masa lampau sebuah perusahaan tidak terkena pelanggaran data, maka di masa depan pun pelanggaran data diantisipasi tidak akan terjadi, dan apabila sebuah perusahaan terkena pelanggaran data di masa lampau, kemungkinan di masa depan pelanggaran data akan terjadi pada perusahaan tersebut.

DAFTAR PUSTAKA

Aas, K., Czado, C., Frigessi, A., dan Bakken, H. 2009. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics* Vol. 44, No. 2, Hal: 182-198.

Bedford, T. dan Cooke, R.M. 2002. Vines: A New Graphical Model for Dependent Random Variables. *The Annals of Statistics* Vol. 33, No. 4, Hal: 1031-1068.

Czado, C. 2010. Pair-Copula Constructions of Multivariate Copula. *Proceedings on Workshop of Lecture Notes in Statistics Vol. 198, Copula Theory and Its Applications*, University of Warsaw: 25-26 September 2009.

Czado, C. dan Nagler, T. 2022. Vine Copula Based Modeling. *Annual Review of Statistics and Its Application* Vol. 9, No. 1, Hal: 453-477.

Durante, F. dan Sempi, C. 2010. Copula Theory: An Introduction. *Proceedings on Workshop of Lecture Notes in Statistics Vol. 198, Copula Theory and Its Applications*, University of Warsaw: 25-26 September 2009.

Fang, Z., Xu, M., Xu, S., dan Hu, T. 2021. A framework for Predicting Data Breach Risk: Leveraging Dependence to Cope with Sparsity. *IEEE Transactions on Information Forensics and Security* Vol. 13, Hal: 2186-2201.

Friederichs, P. 2007. An introduction to extreme value theory. *COPS Summer School*.

Herawati, N., Nisa, K., dan Setiawan, E. 2017. The Optimal Bandwidth for Kernel Density Estimation ff Skewed Distribution: A Case Study on Survival Time Data of Cancer Patients. *Prosiding Seminar Nasional Metode Kuantitatif 2017* Vol. 1, No. 1, Hal: 380-388. Jurusan Matematika FMIPA Universitas Lampung.

Hu, Y. 2013. Extreme Value Mixture Modelling with Simulation Study and Applications in Finance and Insurance. *Tesis*. Department of Mathematics and Statistics Canterbury University New Zealand.

Identity Theft Resource Center. 2021. *ITRC's Notified - The ITRC's Convenient, Comprehensive, Source for Data Breach Information*. Tersedia: https://notified.idtheftcenter.org/s/ (diakses pada tanggal 20 Desember 2021).

Kang, S. dan Song, J. 2017. Parameter and quantile estimation for the generalized Pareto distribution in peaks over threshold framework. *Journal of the Korean Statistical Society* Vol. 46, No. 4, Hal: 487-501.

MacDonald, A.E., *et al*. 2011. A Flexible Extreme Value Mixture Model. *Computational Statistics and Data Analysis* Vol. 55, No. 6, Hal: 2137-2157.

Nelsen, R.B. 2006. *An Introduction to Copula 2nd Ed*. New York: Springer.

Sun, H., Xu, M., dan Zhao, P. 2020. Modeling Malicious Hacking Data Breach Risks. *North American Actuarial Journal* Vol. 25, No. 4, Hal: 484-502.

U.S. Department of Health and Human Services. 2022. *Breach Portal: Notice to the Secretary of HHS Breach of Unsecured Protected*. Tersedia: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf. (diakses pada tanggal 11 Januari 2022).

U.S. Department of Health and Human Services Administration for Children and Families. 2015. *Information Memorandum: Information Security Programs and Guidelines for Responding to Data Breaches*. Tersedia: https://www.acf.hhs.gov/sites/default/files/documents/cb/ im1504.pdf (diakses pada tanggal 31 Mei 2022).