Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Inna Valdikus

Titre du mémoire : Méthodes de prédiction de la sinistralité et analyse des dépendances du risque cyber

Confidentialité : ☐ NON ☑ OUI (Durée : ☑ 1 an ☐ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus.

Membres présents du jury de la filière : Signature :

Entreprise :

Nom : FORSIDES

Signature :

Directeur de mémoire en entreprise

Membres présents du jury de l'Institut des Actuaires : Signature :

Nom : Thomas Carlier

Signature :

Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise :

Signature du candidat :

Inna VALDIKUS

Mémoire de fin d'étude

ISUP

Méthodes de prédiction de la sinistralité et analyse des dépendances du risque cyber

Table des matières

Remerciements

Je tiens en premier lieu à remercier Thomas Carlier, mon tuteur de stage pour son implication et l'ensemble de ses conseils tout au long de la réalisation et de la rédaction de ce mémoire.

Je remercie Arnaud Cohen de m'avoir accueilli au sein de son cabinet pour mon stage et Pascale Quennelle qui m'a donné l'opportunité de réaliser ma première mission client.

Mes remerciements s'adressent également à Anne-Sophie Molines, Najib Aissaoui et de nouveau à Pascale Quennelle pour m'avoir encadré et aidé lors de mes missions.

Je remercie tous ceux qui ont pu contribuer d'une façon ou d'une autre à ce mémoire ainsi que toute l'équipe de FORSIDES pour leur accueil et leur bonne humeur.

J'aimerais aussi remercier Olivier Lopez pour ses conseils et ses remarques.

Je souhaite remercier également ma famille pour m'avoir soutenue tout au long de ma scolarité.

Résumé

Le risque cyber est un risque émergent qui se place à la première place des risques les plus redoutés par les entreprises. Le nombre d'incidents cyber a augmenté de façon exceptionnelle depuis des années, mais les coûts pour l'économie restent difficiles à estimer. Le potentiel dévastateur du risque cyber peut être similaire à celui des catastrophes naturelles, mais la réponse apportée par les assureurs reste encore timide.

Dans le contexte actuel d'absence de données des sinistres, le risque cyber reste l'un des risques le plus difficile à chiffrer et à modéliser. Cependant les conséquences d'un incident cyber peuvent être dramatiques.

Les assureurs restent réticents tant que la mutualisation n'est pas assurée. Le réel enjeu est donc la sensibilisation des entreprises au risque cyber. Faute de mutualisation et d'historique de données disponibles très limité, les assureurs désespèrent de trouver un modèle économique viable. Les conséquences qui en découlent sont une hausse des taux de primes et des franchises et une baisse des indemnisations proposées.

L'objectif de ce mémoire est d'essayer d'approfondir les connaissances de ce risque en s'appuyant sur les données de la table PRC et d'étudier des modèles de prédiction de la sinistralité cyber ainsi que la dépendance entre des attaques, et ce afin d'apporter une meilleure compréhension du risque dans le cadre assurantiel.

La première partie de ce mémoire a pour objet de présenter le contexte de l'étude, le risque cyber et les problématiques posées par ce risque émergent.

La deuxième partie est consacrée à la présentation de la base des données utilisée dans le cadre de ce mémoire, suivie par l'analyse comparative des méthodes de prédiction de la gravité des sinistres issues des Séries Temporelles et des méthodes du Machine Learning.

La troisième partie est consacrée à l'étude des dépendances observées entre les sinistres et leur modélisation par les méthodes de copules.

Enfin une application dans le cadre assurantiel est présentée dans la dernière partie. Nous y aborderons une analyse de tarification de garantie contre la violation de données pour des entreprises, en se basant sur des données simulées à partir de la base PRC. Cette partie mettra en application les modèles de prédiction de la gravité et du nombre de sinistres vus dans la partie 2, s'appuyant également sur des modèles issus de l'assurance paramétrique.

Abstract

Cyber risk is an emerging risk that ranks first among the most feared by companies. The number of cyber incidents has been rising sharply for years, but the costs to the economy remain difficult to estimate. The devastating potential of cyber risk can be similar to that of natural disasters, but the response from insurers is still timid.

In the current context of lack of claims data, cyber risk remains one of the most difficult risks to quantify and model. However, the consequences of a cyber incident can be dramatic.

Insurers remain reluctant until mutualization is assured. The real challenge is to make companies aware of the cyber risk. Due to the lack of pooling and the very limited historical data available, insurers are desperate to find a viable business model. The consequences are higher premium rates, higher deductibles and lower compensation.

The objective of this paper is to try to deepen the knowledge of this risk based on the data of the PRC table and to study predictive models of the cyber loss experience as well as the dependency structure of attacks, in order to bring a better understanding of the risk in the insurance framework.

The first part of this paper presents the context of the study, the cyber risk and the issues raised by this emerging risk.

The second part is devoted to the presentation of the database used in this work, followed by the comparative analysis of the methods of prediction of the severity of the losses resulting from the Time Series and the Machine Learning methods.

The third part is devoted to the study of the observed dependencies between claims and their modeling by copula methods.

Finally, an application in the insurance context is presented in the last part.

Note de Synthèse

Le risque cyber est un risque émergent qui se place à la première place des risques les plus redoutés par les entreprises. Le nombre d'incidents cyber a augmenté de façon exceptionnelle depuis des années, mais les coûts pour l'économie restent difficiles à estimer. Le potentiel dévastateur du risque cyber peut être similaire à celui des catastrophes naturelles, mais la réponse apportée par les assureurs reste encore timide.

Dans le contexte actuel d'absence de données des sinistres, le risque cyber reste l'un des risques le plus difficile à chiffrer et à modéliser. Cependant les conséquences d'un incident cyber peuvent être dramatiques.

Les assureurs restent réticents tant que la mutualisation n'est pas assurée. Le réel enjeu est donc la sensibilisation des entreprises au risque cyber. Faute de mutualisation et d'historique de données disponibles très limité, les assureurs désespèrent de trouver un modèle économique viable. Les conséquences qui en découlent sont une hausse des taux de primes et des franchises et une baisse des indemnisations proposées.

L'objectif de ce mémoire est d'essayer d'approfondir les connaissances de ce risque en s'appuyant sur les données de la table PRC et d'étudier des modèles de prédiction de la sinistralité cyber ainsi que la dépendance entre les attaques, et ce afin d'apporter une meilleure compréhension du risque dans le cadre assurantiel.

La première partie de ce mémoire a pour objet de présenter le contexte de l'étude, le risque cyber et les problématiques posées par ce risque émergent.

La deuxième partie est consacrée à la présentation de la base des données utilisée dans le cadre de ce mémoire, suivie par l'analyse des méthodes de prédiction de la gravité des sinistres issues des Séries Temporelles et des méthodes du Machine Learning.

Ce mémoire ne traite que le cas de violation de données.

La troisième partie est consacrée à l'étude des dépendances observées entre les sinistres et leur modélisation par les méthodes de copules.

Enfin, une application dans le cadre assurantiel est présentée dans la dernière partie. Nous y aborderons une analyse de tarification de garantie contre la violation de données pour des entreprises, en se basant sur des données simulées à partir de la base PRC. Cette partie mettra en application les modèles de prédiction de la gravité et du nombre de sinistres vus dans la partie 2, s'appuyant également sur des modèles issus de l'assurance paramétrique.

Prédiction de la sinistralité du risque cyber

Le prétraitement de la base de données PRC

La base de données PRC est l'une des plus grandes bases de données répertoriant des violations de données disponibles publiquement depuis 2005 et qui a été largement étudiée dans la littérature.

Cette base contient 8927 enregistrements. Les fuites de données recensées ont été rapportées par des agences gouvernementales ou via des sources médiatiques. La base PRC comporte 11 variables, dont 2 variables spatiales (Longitude, Latitude), une variable indiquant la date de notification de l'incident et 8 variables qualitatives, excepté la variable cible indiquant le nombre d'enregistrements compromis par une fuite.

La base PRC ne dispose pas de coût financier résultant d'un incident cyber.

Cependant, le nombre d'enregistrements est un élément clé pour pouvoir mesurer la sévérité de l'incident. La variable cible (Total Records) est très volatile, ses valeurs se répartissent entre 0 et 3 milliards.

Dans la base PRC plus de la moitié des entreprises attaquées font partie du secteur médical. Elle présente un certain biais, car même si le secteur médical fait partie des secteurs les plus touchés par le risque cyber, son importance dans la base depuis 2010 peut être due à l'obligation de notifier les incidents cyber dans le domaine médical à partir d'un certain niveau de gravité de fuite.

La base PRC reste limitée, il est évident qu'elle ne reflète pas la situation réelle concernant le nombre de fuites observées : il y a certainement plus d'incidents réellement survenus que d'incidents déclarés ; les entreprises étant réticentes à déclarer des fuites des données pour éviter une mauvaise publicité.

Méthodes de classification

Pour pouvoir utiliser la majorité des méthodes du Machine Learning, la variable cible représentant le nombre d'enregistrements compromis par un incident cyber a été recodée en 4 modalités de la manière suivante :

Modalité	Range
XS	[1,500]
S	[501,10000]
M	[10001,100000]
L	>100000

Cette section se concentre sur les algorithmes d'apprentissage automatique ML utilisés dans l'objectif de trouver l'algorithme le plus performant pour la prédiction de la gravité d'un sinistre cyber.

Les modèles de Machine Learning testés sont les suivants :

- Les arbres de décision,
- Les forets aléatoires,
- Le naïf bayésien,

- La régression logistique multinomiale,
- Le gradient boosting,
- Les réseaux de neurones.

Pour éviter le surapprentissage, le jeu de données est séparé en deux échantillons distincts : un ensemble d'apprentissage (train) et un ensemble de test. L'échantillon de test est obtenu par une sélection aléatoire des lignes de la base de données totale.

Les résultats obtenus sont les suivants :

Modèles	Accuracy	Kappa
Arbres de décision	0.661	0.327
Random Forest	0.669	0.323
XGBoost	0.671	0.329
Multinomial logistic regression	0.677	0.323
Naive Bayesian	0.681	0.342
Neural Network	0.705	0.351

Dans le cas de l'analyse par les méthodes du Machine Learning, tous les modèles rencontrent le même problème pour la prédiction des sinistres extrêmes. Les modalités de la variable représentant la gravité du sinistre sont déséquilibrées. La classe S est surreprésentée en nombre. L'utilisation de l'accuracy n'est pas la plus adaptée pour ce type de données. La solution possible pour le problème des données déséquilibrées sont des techniques de sous ou sur-échantillonnage. En considérant l'indicateur ROC AUC, les modèles qui sortent un peu du lot sont Random forest et XGboost. Cependant le modèle le plus performant reste le modèle de réseaux de neurones, d'autant que ce modèle-ci n'est pas autant concerné par le problème de données déséquilibrées, car il n'utilise que la variable textuelle (description de l'incident cyber) dans la prédiction de la gravité de l'incident cyber.

Méthodes de régression

Cette section se concentre sur les algorithmes de Machine Learning utilisés à des fins de régression dans l'objectif de trouver l'algorithme le plus performant pour la prédiction de la gravité d'un sinistre cyber log transformée et aussi pour la prédiction du nombre d'incidents survenus.

Les modèles de régression testés initialement ne donnant pas de résultats probants ($R2<0.05$), nous avons donc procédé à un retraitement de la base pour pouvoir utiliser dans les modèles de régression les données mensuelles en fonction du secteur et de la zone géographique. Ce procédé a permis d'améliorer sensiblement les performances des modèles. Même les modèles les moins performants donnent un R2 supérieur à 0.58.

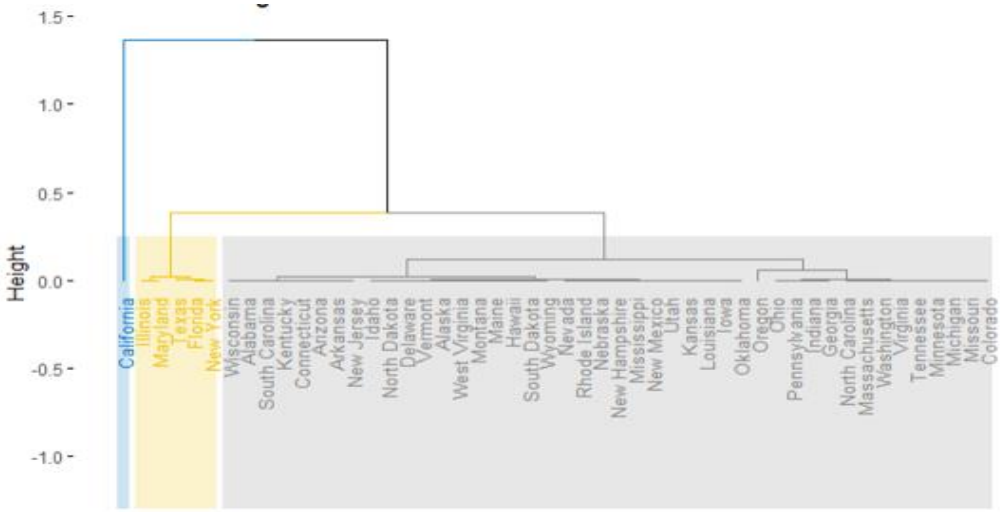
Comme pour la classification, les résultats de tous les modèles seront comparés afin de sélectionner le modèle le plus performant au sens des critères de validation choisi : RMSE.

La base PRC est caractérisée par une forte hétérogénéité spatiale, c'est pourquoi le regroupement par zone de risque homogène s'est avéré essentiel.

Le regroupement de ces zones géographiques sera réalisée par les méthodes d'analyse de données factorielles.

- *ACP : analyse en composantes principales suivie de*
- *CAH : classification ascendante hiérarchique*

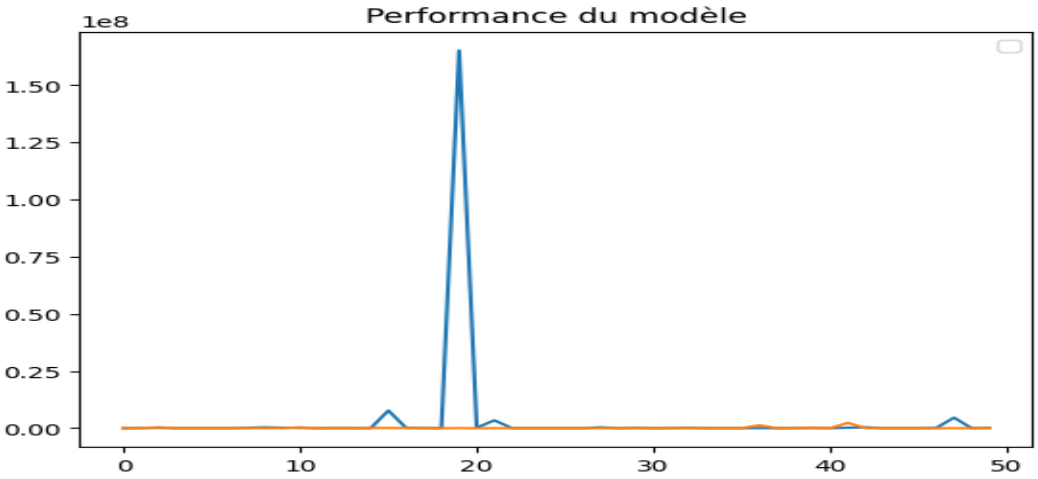
La CAH est une méthode de classification faisant partie des méthodes non supervisées. Elle a pour objectif de construire une hiérarchie sur les individus et se présente sous la forme d'un arbre permettant de visualiser les distances entre individus et groupes d'individus (dendrogramme). La CAH nécessite de choisir deux paramètres.



Le découpage obtenu est le suivant :

- Cluster 3: California
- Cluster 2: Florida, Texas, Maryland, New York; Illinois.
- Cluster 1: Tous les états restants

Une transformation logarithmique a été effectuée sur la variable cible (nombre d'enregistrements compromis mensuel moyen) car une grande variabilité existe : les valeurs sont réparties entre 0 et 3 milliards et les études antérieures sur la variable non transformée donnaient des résultats peu probants : la métrique RMSE était particulièrement élevée (RMSE : 83 255 420). Comme nous pouvons observer sur le graphique, la prédiction est très basse :



Nous avons aussi cherché à prédire le log transformé de la moyenne de la sévérité mensuelle des sinistres cyber et le nombre d'incidents par mois (fréquence).

Pour avoir une meilleure vue d'ensemble des résultats, un tableau résumant les performances des modèles de régression est présenté :

Modèle	NB d'incidents		Log(gravité)	
	Naïf	Grid search	Naif	Grid search
Arbres de décision	4.184	3.152	1.590	1.490
Random Forest	3.727	3.314	1.577	1.528
Gradient Boosting	3.359	3.196	1.567	1.484
Réseaux de neurones	2.540	2.496	1.569	1.545

Le modèle de réseaux de neurones s'est révélé particulièrement performant dans la prédiction du nombre d'incidents mensuels (à savoir pour l'estimation de la fréquence : nombre de sinistres). Concernant la sévérité des incidents, tous les modèles ont montré des performances relativement proches, cependant c'est le modèle de gradient boosting qui s'est montré le plus performant.

Nous considérons que la performance des modèles de neurones peut être significativement améliorée en déterminant une meilleure calibration. Cependant en gagnant en précision, on « gagne » également en complexité, car il s'agît d'une boite noire.

Les autres modèles testés présentent l'avantage d'être plus simples et plus rapides à calibrer pour des performances sensiblement équivalentes.

Séries temporelles

Comme pour les modèles de régression, la classification selon les zones de risque homogène face au risque cyber s'avère donc essentielle. Dans un premier temps, certains modèles théoriques sont présentés.

L'objectif de cette partie est de développer plusieurs modèles afin d'estimer :

- Prédiction de la sévérité des attaques (nombre moyen d'enregistrements compromis pour l'année N+1 en utilisant les données mensuelles moyennes de la sinistralité) :
- Prédiction de la fréquence : nombre de sinistres pour l'année N+1 en utilisant les données mensuelles

La transformation logarithmique de la variable (nombre d'enregistrements compromis) évoquée plus haut est là encore essentielle en raison de la grande variabilité des valeurs (réparties entre 0 et 3 milliards).

L'analyse d'une série entière est réalisée avant de considérer chaque zone géographique séparément. Le modèle le plus performant pour la prédiction de la fréquence (nombre d'incidents mensuels) et de la gravité, avant de faire la distinction par zones géographiques, est ARIMA.

Les modèles de séries temporelles seraient moins performants que les modèles de Machine Learning tant pour la prédiction de la gravité d'incident que pour la fréquence.

- Sévérité des incidents (nombre d'enregistrements violés)

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA (3)
RMSE	2.52	2.47	2.55	3.15	2.66	2.77	2.49

- Fréquence (nombre d'incidents)

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	18.62	12.89	21.53	15.77	19.73	17.50

Pour améliorer les performances des modèles des séries temporelles, la décomposition par zone géographique est effectuée.

- Fréquence

Zone géographique 1 (nombre de sinistres par mois) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	9.90	11.32	12.46	6.22	10.16	9.80

Zone géographique 2 (nombre de sinistres par mois) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	3.79	4.95	3.62	6.61	3.48	4.11

Zone géographique 3 (nombre de sinistres par mois) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	2.14	2.18	2.33	2.49	2.08	2.18

- Sévérité

Zone géographique 1 (nombre d'enregistrements moyen par mois) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA (3)
RMSE	1.97	2.31	2.83	3.12	2.77	2.76	2.22

Zone géographique 2 (nombre d'enregistrements moyen par mois) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA
RMSE	2.31	2.14	2.47	2.48	2.31	2.25	2.39

Zone géographique 3 (nombre d'enregistrements moyen par mois) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA(3)
RMSE	2.01	3.44	4.30	4.63	4.34	4.34	3.33

Le regroupement par zone géographique a un effet bénéfique sur les performances prédictives des modèles des séries temporelles, tout particulièrement pour la prédiction de la fréquence.

En synthèse, nous avons observé dans ce mémoire que les méthodes de Machine Learning se sont avérées plus performantes pour la prédiction de la sinistralité cyber que les méthodes de séries temporelles.

L'étude des dépendances

Dans cette partie, la structure de dépendance de la gravité des pertes mensuelles moyennes de différents types d'incidents cyber sera analysée. On a trois types de dépendances possibles :

- Les pertes par zone géographique
- Les pertes intersectorielles (secteur médical, financier, commerce ou administration publique),
- Les pertes par type de violation subie (piratage, perte d'un appareil électronique, divulgation par une personne interne à l'entreprise, perte des documents papier).

Comme dans le cas des séries temporelles, on a procédé à un regroupement des sinistres par zone de risque homogène.

Pour l'analyse des dépendances, nous nous sommes basés sur la construction de copules. Une copule est une fonction de répartition multidimensionnelle ayant des lois marginales uniformes sur [0,1]. Au préalable, il faut trouver les lois marginales des distributions pour chaque type de risque. Le choix de la copule dépend du risque que nous voulons modéliser. Pour un risque présentant des valeurs extrêmes, il est logique de s'intéresser aux copules extrêmes.

La méthode des copules présente plusieurs avantages : elle permet de donner une représentation de la dépendance entre les risques mais aussi de décrire un comportement de chaque risque (loi marginale).

Application au domaine de l'assurance

Il nous a semblé intéressant de nous pencher sur un exemple de tarification d'une garantie contre la violation de données à destination d'entreprises.

Pour ce faire, nous avons entrepris la construction d'un portefeuille d'assurés par le biais de la base PRC, en répliquant dans un premier temps un portefeuille de sinistres fictifs sur 10 années, de 2013 à 2022.

Sur la base des données reconstituées, nous avons appliqué les modèles les plus performants de nos premières analyses afin d'estimer la sinistralité à venir en année N+1.

Dans notre modèle coût-fréquence, la fréquence des sinistres est un paramètre d'étude, dont nous analyserons l'impact sur le tarif. Quant au coût, nous nous sommes essentiellement basés sur les modèles d'assurance paramétriques de Jacobs et de Farkas.

L'objet de cette partie est d'analyser sur la base d'un portefeuille de violation de données inspiré de la table PRC, la sensibilité d'un tarif de prime pure d'assurance aux variables et facteurs suivants :

- Tarif segmenté vs tarif global

- Impact d'une franchise
- Impact d'un plafond d'indemnisation
- Sensibilité à la fréquence des sinistres

Le tableau récapitulatif suivant montre les sensibilités du tarif aux différents paramètres testés.

Jeu de fréquence 1

			Global	MED	ADM	RETAIL	BSF
		Fréquence	***18%***	***20%***	***15%***	***15%***	***10%***
		Tarif	**484 230**	**386 501**	**791 439**	**515 523**	**364 868**
Sensibilité	Franchise	Franchise 100K	-4%	-5%	-2%	-3%	-3%
		Franchise 300K	-11%	-15%	-6%	-8%	-7%
		Franchise 500K	-18%	-25%	-9%	-13%	-12%
	Plafond	Plafond 2M	-50%	-37%	-67%	-57%	-59%
		Plafond 5M	-21%	-13%	-36%	-22%	-24%
		Plafond 8M	-10%	-5%	-18%	-7%	-11%
	Franchise + Plafond	F 100K - P 2M	-53%	-42%	-69%	-60%	-61%
		F 300K - P 5M	-32%	-28%	-41%	-30%	-31%
		F 500K - P 8M	-27%	-30%	-27%	-20%	-22%

Les analyses réalisées et développées dans ce mémoire mettent en évidence plusieurs éléments importants :

- Les tarifs restent très importants (tarif global environ à 500K€).
- L'effet de mutualisation intersectoriel s'avère très bénéfique pour les entreprises / organismes du secteur administratif et dans une moindre mesure pour celles du retail, au détriment des entreprises du secteur médical et du secteur bancaire et des services financiers.
- Comme attendu, la mise en place de plafonds relativement bas entraine la baisse tarifaire la plus marquée. Elle présente malgré tout le désavantage d'un reste à charge important en cas de sinistre conséquent pour les entreprises touchées ; en contrepartie, les entreprises subissant de faibles sinistres ou épargnées seraient largement avantagées par ce type de solution.

En synthèse, ces analyses confirment les difficultés rencontrées pour la mise en place de garanties contre la violation de données, et plus particulièrement la difficulté de mobilisation des TPE et PME, pour lesquelles les tarifs des polices d'assurance demeurent très élevés.

L'enjeu autour de l'assurance cyber est donc particulièrement important, au regard notamment du niveau de couverture actuel d'entreprises qui semblent en moyenne faiblement prémunies contre les impacts d'éventuels incidents. L'une des raisons derrière ces observations réside dans le fait que l'estimation du risque cyber s'appuie sur un faible volume de données, rendant le risque cyber particulièrement difficile à évaluer.

Executive Summary

Cyber risk is an emerging risk that ranks first among the most feared risks for businesses. The number of cyber incidents has grown exponentially over the years, but the overall cost to the economy as a whole remains difficult to estimate. The devastating potential of cyber risk may be similar to that of natural disasters, but the response from insurers is still tentative.

In the current context of lack of claims data, cyber risk remains one of the most difficult risks to quantify and model. However, the consequences of a cyber incident can be dramatic.

Insurers remain reluctant until mutualization is assured. The real challenge is to make companies aware of the cyber risk. Due to the lack of pooling and the very limited historical data available, insurers are desperate to find a viable business model. The consequences are higher premium rates, higher deductibles and lower compensation.

The objective of this paper is to try to deepen the knowledge of this risk by using the PRC table data and to study predictive models of cyber claims and the dependency between attacks, in order to bring a better understanding of the risk in the insurance framework.

The first part of this paper presents the context of the study, the cyber risk and the issues raised by this emerging risk.

The second part is devoted to the presentation of the database used in this thesis, followed by the analysis of the methods of prediction of the severity of the losses resulting from Time Series and Machine Learning methods.

This dissertation only deals with the case of data breach.

The third part is devoted to the study of the observed dependencies between claims and their modeling by copula methods.

Finally, an application in the insurance context is presented in the last part. We will discuss a pricing analysis of data breach coverage for companies, based on simulated data from the PRC database. This section will apply the severity and loss prediction models seen in Section 2, also based on models from parametric insurance.

Predicting cyber risk losses

Preprocessing of the PRC database

The PRC database is one of the largest databases of publicly available data breaches since 2005 and has been widely studied in the literature.

The database contains 8927 records. The data leaks were reported by government agencies or via media sources. The PRC database has 11 variables, including two spatial variables (Longitude, Latitude), one variable indicating the date of notification of the incident, and eight categorical variables, except for the target variable indicating the number of records compromised by a leak.

The PRC database does not have a financial cost resulting from a cyber incident.

However, the number of records is a key element to be able to measure the severity of the incident. The target variable (Total Records) is very volatile, with values ranging from 0 to 3 billion.

In the PRC database, more than half of the companies attacked are in the medical sector. It presents a certain bias, because even if the medical sector is one of the sectors most affected by cyber risk, its importance in the database since 2010 may be due to the obligation to notify cyber incidents in the medical field from a certain level of leak severity.

The PRC database remains limited, it is obvious that it does not reflect the real situation regarding the number of leaks observed: there are certainly more incidents that actually occurred than reported; companies being reluctant to report data leaks to avoid bad publicity.

Classification methods

To be able to use most of the Machine Learning methods, the target variable representing the number of records compromised by a cyber incident has been recoded into 4 modalities as follows:

Modalité	Range
XS	[1,500]
S	[501,10000]
M	[10001,100000]
L	>100000

This section focuses on the ML machine learning algorithms used with the objective of finding the best performing algorithm for predicting the severity of a cyber disaster.

The Machine Learning models tested are:

- Decision Trees,
- Random Forest,
- Naive Bayesian,
- Multinomial logistic regression,
- Gradient boosting
- and neural networks.

To avoid overlearning, the dataset is separated into two distinct samples: a training set (train) and a test set. The test sample is obtained by a random selection of rows from the total database.

The results obtained are as follows:

Modèles	Accuracy	Kappa
Arbres de décision	0.661	0.327
Random Forest	0.669	0.323
XGBoost	0.671	0.329
Multinomial logistic regression	0.677	0.323
Naive Bayesian	0.681	0.342
Neural Network	0.705	0.351

In the case of the Machine Learning analysis, all models have the same problem in predicting big losses. The modalities of the variable representing the severity of the loss are unbalanced. The S class is overrepresented in number. The use of accuracy is not the most suitable for this type of data. The possible solution to the problem of unbalanced data are under- or over-sampling techniques. Considering the ROC AUC indicator, the models that stand out a little are Random forest and XGboost. However, the best performing model remains the neural network model, especially since this model is not as concerned by the problem of unbalanced data, because it only uses the textual variable (description of the cyber incident) in the prediction of the severity of the cyber incident.

Regression Methods

This section focuses on Machine Learning algorithms used for regression purposes with the objective of finding the best performing algorithm for predicting the severity of a cyber log transformed loss and also for predicting the number of incidents that occurred.

The regression models tested initially did not give convincing results ($R2<0.05$), so we reprocessed the database to be able to use monthly data according to sector and geographical area in the regression models. This process allowed us to significantly improve the performance of the models. Even the worst performing models gave an R2 greater than 0.58.

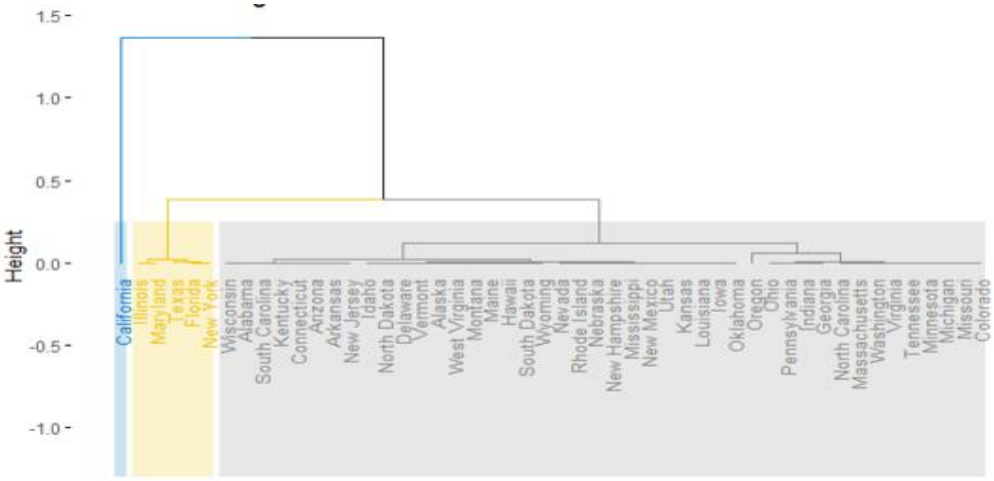
As for the classification, the results of all the models will be compared in order to select the best performing model according to the chosen validation criteria: RMSE.

The PRC database is characterized by a strong spatial heterogeneity, which is why the grouping by homogeneous risk area was essential.

The grouping of these geographical areas will be done by factorial data analysis methods.

- PCA: principal component analysis followed by
- HAC: hierarchical ascending classification

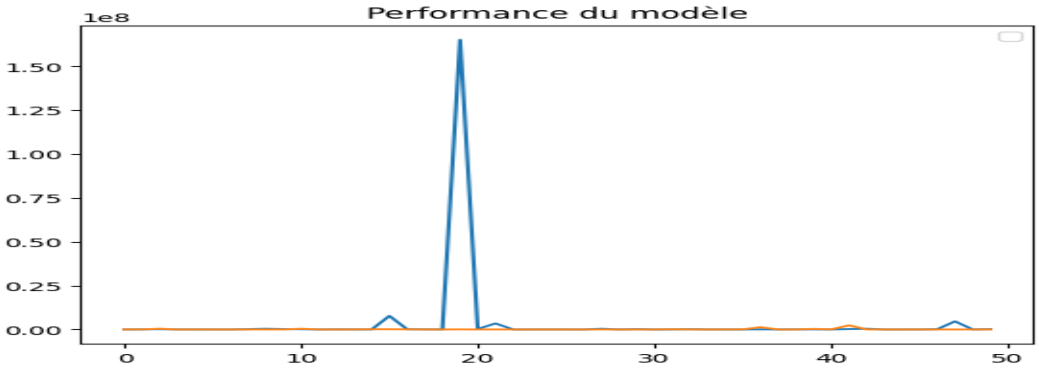
HAC is an unsupervised classification method. Its objective is to build a hierarchy on the individuals and is presented in the form of a tree allowing to visualize the distances between individuals and groups of individuals (dendrogram). The HAC requires the choice of two parameters.



The resulting clustering is as follows:

- Cluster 3: California
- Cluster 2: Florida, Texas, Maryland, New York; Illinois.
- Cluster 1: All remaining states

A log transformation was performed on the target variable (average monthly compromise records) because a large variability exists: the values are distributed between 0 and 3 billion and previous studies on the untransformed variable yielded inconclusive results. The RMSE metric was particularly high (RMSE : 83 255 420). As we can see on the graph, the prediction is very low:



We also sought to predict the transformed log of the average monthly cyber loss severity and the number of incidents per month (frequency).

To get a better overview of the results, a table summarizing the performance of the regression models is presented :

Modèle	NB d'incidents		Log(gravité)	
	Naïf	Grid search	Naif	Grid search
Arbres de décision	4.184	3.152	1.590	1.490
Random Forest	3.727	3.314	1.577	1.528
Gradient Boosting	3.359	3.196	1.567	1.484
Réseaux de neurones	2.540	2.496	1.569	1.545

The neural network model performed particularly well in predicting the number of monthly incidents (i.e. for the estimation of the frequency: number of claims). Concerning the severity of the incidents, all the models showed relatively close performances, however it is the gradient boosting model which showed the best performance.

We consider that the performance of neural models can be significantly improved by determining a better calibration. However, by gaining in precision, we also "gain" in complexity, because it is a black box.

The other models tested have the advantage of being simpler and quicker to calibrate for roughly equivalent performance.

Time series

As with regression models, classification according to zones of homogeneous risk in the face of cyber risk is therefore essential. First, some theoretical models are presented.

The objective of this part is to develop several models to estimate:

- Attack severity prediction (average number of compromised records for year N+1 using average monthly loss data) :

- Frequency prediction: number of claims for year N+1 using monthly data

The logarithmic transformation of the variable (number of compromised records) mentioned above is again essential because of the large variability of the values (distributed between 0 and 3 billions).

The analysis of an entire series is performed before considering each geographical area separately. The best performing model for predicting frequency (number of monthly incidents) and severity, before distinguishing by geographical area, is ARIMA

The time series models would perform less well than the Machine Learning models for the prediction of both incident severity and frequency.

Incident severity (number of records)

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA (3)
RMSE	2.52	2.47	2.55	3.15	2.66	2.76	2.49

Frequency (number of incidents)

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	18.62	12.89	21.53	15.77	19.73	17.50

To improve the performance of the time series models, the decomposition by geographical area is performed.

- Frequency

Geographic Area 1 (number of claims per month) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	9.90	11.32	12.46	6.22	10.16	9.80

Geographic Area 2 (number of claims per month) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	3.79	4.95	3.62	6.61	3.48	4.11

Geographic Area 3 (number of claims per month) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	2.14	2.18	2.33	2.49	2.08	2.18

- Severity

Geographic Area 1 (average number of records per month) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA(3)
RMSE	1.97	2.31	2.83	3.12	2.77	2.76	2.22

Geographic Area 2 (average number of records per month) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA
RMSE	2.31	2.14	2.47	2.48	2.31	2.25	2.39

Geographic Area 3 (average number of records per month) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA(3)
RMSE	2.01	3.44	4.30	4.63	4.34	4.34	3.33

Clustering by geographic area has a beneficial effect on the predictive performance of time series models, especially for frequency prediction. In summary, we observed in this dissertation that Machine Learning methods performed better for cyber claims prediction than time series methods.

The study of dependencies

In this section, the dependency structure of the average monthly loss severity of different types of cyber incidents will be analyzed. We have two possible types of dependencies:

- Losses by geographic area
- Losses across sectors (medical, financial, commercial or public administration),

- Losses by type of breach (hacking, loss of an electronic device, disclosure by an internal person, loss of paper documents).

As in the case of the time series, the losses were grouped by homogeneous risk areas.

For the analysis of dependencies, we have based ourselves on the construction of copulas. A copula is a multidimensional distribution function with uniform marginal laws on [0,1]. Beforehand, we have to find the marginal laws of the distributions for each type of risk. The choice of the copula depends on the risk we want to model. For a risk with extreme values, it is logical to be interested in extreme copulas.

The copula method has several advantages: it allows to give a representation of the dependence between the risks but also to describe a behavior of each risk (marginal law).

Application to the insurance domain

We thought it would be interesting to look at an example of pricing a data breach insurance policy for companies.

To do so, we have undertaken the construction of a portfolio of insureds through the PRC database, by first replicating a portfolio of fictitious claims over 10 years, from 2013 to 2022.

Based on the reconstructed data, we applied the best performing models from our initial analyses to estimate the future claims experience in year N+1.

In our cost-frequency model, the frequency of claims is a study parameter, whose impact on the rate will be analyzed. As for the cost, we have essentially based ourselves on the parametric insurance models of Jacobs and Farkas.

The purpose of this section is to analyze, on the basis of a portfolio of data violations inspired by the PRC table, the sensitivity of a pure insurance premium tariff to the following variables and factors

- Segmented rate vs. Global rate
- Impact of a deductible
- Impact of an indemnity cap
- Sensitivity to loss frequency

The following summary table shows the sensitivities of the rate to the different parameters tested.

Jeu de fréquence 1

			Global	MED	ADM	RETAIL	BSF
		Fréquence	*18%*	*20%*	*15%*	*15%*	*10%*
		Tarif	484 230	386 501	791 439	515 523	364 868
Sensibilité	Franchise	Franchise 100K	-4%	-5%	-2%	-3%	-3%
		Franchise 300K	-11%	-15%	-6%	-8%	-7%
		Franchise 500K	-18%	-25%	-9%	-13%	-12%
	Plafond	Plafond 2M	-50%	-37%	-67%	-57%	-59%
		Plafond 5M	-21%	-13%	-36%	-22%	-24%
		Plafond 8M	-10%	-5%	-18%	-7%	-11%
	Franchise + Plafond	F 100K - P 2M	-53%	-42%	-69%	-60%	-61%
		F 300K - P 5M	-32%	-28%	-41%	-30%	-31%
		F 500K - P 8M	-27%	-30%	-27%	-20%	-22%

The analyses carried out and developed in this report highlight several important elements:

- The rates remain very high (overall rate around 500K€)
- The inter-sectoral mutualization effect is very beneficial for companies/organizations in the administrative sector and to a lesser extent for those in the retail sector, to the detriment of companies in the medical sector and the banking and financial services sector
- As expected, the implementation of relatively low reimbursement ceilings leads to the most significant price decrease. The disadvantage is that the companies affected would have to pay a large amount of money in the event of a major claim ; on the other hand, companies with small claims or that are spared would benefit greatly from this type of solution.

In summary, these analyses confirm the difficulties encountered in setting up data breach coverage, and more particularly the difficulty of mobilizing VSEs and SMEs for which insurance policies remain very high.

The issue of cyber insurance is therefore particularly important, especially in light of the current level of coverage of companies that seem to be poorly protected against the impact of potential incidents.

One of the reasons behind these observations lies in the fact that the estimation of cyber risk is based on a low volume of data, making cyber risk particularly difficult to assess.

Introduction

Dans une économie qui se digitalise de plus en plus, le risque cyber est encore très mal compris. Les entreprises et les collectivités publiques peinent à prendre la mesure du danger alors que les pirates ne cessent de gagner en expertise. Depuis les années quatre-vingt-dix, le nombre et la gravité des incidents cyber ne cessent d'augmenter. Selon le baromètre des risques de 2022 d'Allianz (AGCS[1]), le risque cyber se place à la première place des risques redoutés par les entreprises devant le risque d'interruption d'activité et le risque des catastrophes naturelles. La hausse des attaques par ransomware en est une des explications de ce phénomène. Parmi les personnes interrogées, 57% la considèrent comme la principale menace cyber. Selon le dernier baromètre du Cesin[2] publié en janvier 2022, plus de 50% d'entreprises auraient subi entre une et trois attaques cyber au cours de l'année 2021. Et on ne parle que des attaques réussies… Ces attaques entraînent souvent des conséquences très lourdes, elles sont susceptibles de paralyser le fonctionnement des entreprises et pouvant même mettre en jeu leur survie, ainsi que celle de leurs clients, de leurs fournisseurs.

Les cyberattaques deviennent de plus en plus efficaces et massives, ce ne sont plus des attaques isolées, mais plutôt une cyber criminalité en bandes organisées. Ces organisations criminelles sont attirées par des gains très élevés et une certaine impunité, car les arrestations de hackers étaient auparavant très rares. Selon le cabinet de conseil Wavestone, le retour sur investissement d'une cyberattaque serait entre 200% et 800%. Les cyberattaques ne sont pas que le fait des organisations criminelles, certains Etats défendent aussi leurs intérêts de cette manière. Les pirates sont présents dans tous les pays, mais certains pays font des cyberattaques leur spécialité. Selon Chainanalysis[3], les hackers liés à la Russie auraient extorqué 400 millions de dollars sous forme de cryptomonnaie en utilisant des attaques de type ransomware, c'est-à-dire 74% de l'ensemble d'extorsions réalisées en 2021.

Selon des experts, les attaques ont augmenté dans le monde entier en 2021. Alessandro Profumo, directeur général du constructeur aéronautique Leonardo a annoncé que la cybercriminalité aurait coûté une somme astronomique de plus de 6000 Md$ en 2021.

Plus récemment, la situation géopolitique et diplomatique mondiale dégradée depuis le début de l'année 2022 a amplifié le nombre d'attaques et leur intensité. Selon un rapport de Microsoft publié le 27/04/2022, 237 cyberattaques auraient été effectuées par des groupes liés à l'Etat russe contre l'Ukraine et ses infrastructures juste avant le début de l'invasion. Ces attaques seraient en préparation depuis le mois de mars 2021. Les soutiens de l'Ukraine seraient aussi des cibles privilégiées. Les tensions politiques actuelles observées nous amènent à anticiper des cyberattaques et de s'y préparer afin d'en limiter les effets.

Dans le contexte actuel d'absence de données des sinistres, le risque cyber reste l'un des risques le plus difficile à chiffrer et à modéliser. Cependant les conséquences d'un incident cyber peuvent être dramatiques tant pour les entreprises que pour les personnes. Par exemple, un

---

1 *AGCS: Allianz Global Corporate & Specialt*
L'enquête annuelle d'Allianz Global Corporate & Specialty (AGCS) analyse les opinions de 2 650 experts, notamment des directeurs généraux, gestionnaires de risques, courtiers et assureurs, dans 89 pays et territoires.
2 *7ème édition du baromètre annuel du CESIN : Enquête exclusive sur la cybersécurité des entreprises françaises*
3 *Chainanalysis : société américaine d'analyse de blockchains spécialisée dans la cybersécurité*

hacker pourrait prendre contrôle d'une voiture connectée. La sécurité nationale et la vie des citoyens pourraient être mises en péril par un piratage de données hautement sensibles.

Le nombre d'incidents cyber a augmenté de façon exceptionnelle depuis des années, mais les coûts pour l'économie restent difficiles à estimer. Le potentiel dévastateur du risque cyber peut être similaire à celui des catastrophes naturelles, mais la réponse apportée par les assureurs reste encore timide.

CyRim[4], une plateforme de recherche a fait une hypothèse d'une attaque cyber de grande ampleur de type ransomware (Bashe attack). Les pertes économiques estimées pourraient atteindre des montants exorbitants compris entre 85 Md$ et 193 Md$.

Une prise de conscience du risque cyber est donc essentielle. Toute entreprise est susceptible d'être touchée par ce risque. Il ne faut plus être un grand groupe pour être dans le viseur des hackers. Les chiffres sont parlants. Avec une augmentation de 255% en 2020, la France est le troisième pays le plus touché par le ransomware et ce sont les TPE et les PME qui sont les principales victimes. Le ransomware est un type d'attaque informatique qui bloque l'accès à l'appareil ou aux fichiers d'une victime en chiffrant les données et qui exige le paiement d'une rançon en échange du rétablissement de l'accès. Vingt pour cent des dirigeants des petites et moyennes entreprises ont déclaré avoir subi au moins une attaque ou tentative de cyber attaques au cours de l'année 2021.

Afin de se prémunir contre les effets potentiellement dévastateurs des attaques cyber, les entreprises ou organismes potentiellement ciblés doivent avant tout mettre en place une politique de prévention des attaques efficace.

Celle-ci passe notamment pour une prise de conscience des dirigeants qui doivent investir dans la prévention en sécurisant plus les systèmes informatiques sur lesquels leurs organisations reposent, mais il est tout aussi primordial d'investir dans l'humain. En effet, certaines études démontrent qu'une grande partie des attaques cyber subies à ce jour le sont en raison d'erreurs humaines, qui facilitent la tâche des hackers.

- **Prévention des attaques** : Identification des vulnérabilités suivie par sécurisation des systèmes d'information ; formation des employés aux enjeux de la sécurité numérique et aux techniques des pirates. L'humain reste le maillon faible de la cybersécurité : la moitié des attaques découlent d'erreurs humaines, c'est pourquoi la formation aux bonnes pratiques en ligne et la sensibilisation à la sécurité du numérique semblent essentielles, puisque 80% des attaques auraient pu être évitées par l'application de mesures simples de sécurité.

Enfin, l'autre moyen de se protéger face au risque cyber, à défaut de s'en prémunir en amont grâce à la prévention, est de se protéger de ses conséquences potentiellement systémiques en s'assurant.

4 *CyRim (cyber risk management) est une plateforme de recherche à laquelle participent des assureurs tels que Scor, Trans Re, le Lloyd's, Aon et MSIG*

- **Le recours à l'assurance** : L'assurance permet de financer la gestion de crise lors d'un incident : appel aux experts de cybersécurité pour limiter les dégâts, indemnisation des dommages financiers liés à l'attaque (perte d'exploitation, frais d'avocats, ou encore frais d'experts en informatique).

Malgré ça, les acteurs de l'économie que sont les grands groupes, les petites et moyennes entreprises (PME) et les entreprises de taille intermédiaire (ETI) semblent ne pas avoir suffisamment pris conscience de la menace que le risque cyber fait peser sur leurs activités et peinent à mettre en place des mesures de prévention ou de gestion de crise et sont également très faiblement assurées contre ce type de risque.

Sur ce dernier point, et selon le rapport LUCY de l'AMRAE[5], sur 140 000 PME réalisant entre 10 et 15 M€ de chiffre d'affaires Il n'y a que 362 entreprises qui ont souscrit une assurance cyber auprès de leur courtier en 2020 et 322 en 2021. Même si cette statistique (taux de couverture de 0,0026% en 2020) peut paraitre biaisée car elle n'inclut pas les assurances souscrites auprès d'agents généraux, il en ressort que le taux de couverture des PME est très faible, pour ne pas dire insignifiant. Plusieurs facteurs peuvent expliquer cette réticence à s'assurer : le manque de temps, de moyens financiers et « l'ignorance de risque ». Les collectivités territoriales se trouvent dans la même situation. Cette sous-couverture fait courir des risques aux entreprises et aux collectivités publiques concernées, mais aussi à l'ensemble de leurs partenaires économiques, fournisseurs, clients et sous-traitants principalement.

Concernant les ETI, le taux de couverture est supérieur (9% en 2021) mais reste malgré tout très faible, démontrant un vrai manque en matière de protection face au risque cyber pour ce type d'entreprise.

Il existe toutefois une inégalité dans la couverture entre les PME et ETI et les grandes entreprises, dont le taux de couverture cyber est de 87% en 2020, même si ce taux a subi une légère baisse en 2021 en passant à 84%.

Cependant, si le taux de couverture des grandes entreprises est élevé et donne l'impression qu'elles sont de fait bien protégées en cas d'adversité, c'est le niveau de leur protection qui semble inadapté et faible au regard des risques d'aujourd'hui.

En effet, l'enquête LUCY, réalisée par L'AMRAE en partenariat avec des grands courtiers spécialisés en risque d'entreprise, met en lumière le fait que malgré une certaine augmentation du volume de primes en 2020, qui est passé de 87 à 130 millions d'euros (+49%) en 2020, celle-ci reste insuffisante par rapport à la hausse du montant des indemnisations qui est passé de de 73 à 217 millions d'euros. Le ratio S/P (Sinistres sur Primes) a donc évolué considérablement en passant de 84% à 167% en une année. En 2021 le ratio S/P s'est amélioré en retrouvant la valeur de 2019 soit 88%. Le marché semble plus rentable qu'en 2020, mais il reste toujours

5 *AMRAE : L'Association pour le Management des Risques et des Assurances de l'Entreprise est l'association professionnelle de référence des métiers du risque et des assurances en entreprise. Elle rassemble plus de 1 500 membres appartenant à plus de 750 organisations privées ou publiques. Elle a publié en 2021 la première étude LUCY (LUmière sur la CYberassurance) objective et exhaustive sur le risque cyber et sa couverture assurantielle. « Cette étude a été menée auprès de huit grands courtiers spécialistes du risque d'entreprise : AON, Diot, Filhet Allard, Marsh, Siaci saint Honoré, Verlingue, Verspieren, Gras Savoye-Willis Towers Watson. Planète CSCA, le syndicat du courtage, a également été mis à contribution, notamment pour avoir une meilleure vision du marché des PME. »*

volatile. Les coûts d'indemnisations ont baissé en 2021 en passant de 217 millions d'euros à 185 millions d'euros, le marché est devenu plus particulièrement rentable sur le segment des grandes entreprises, La situation était d'autant plus critique puisque l'étude de 2020 montrait que le ratio S/P a plus que quadruplé, passant de 44 % à 190 % en 2020.

Cette augmentation était surtout due aux quatre sinistres de très forte intensité déclarés par des grandes entreprises. Ils ne représentent que 1% des sinistres déclarés, chaque sinistre a coûté aux assureurs entre 10 et 40 millions d'euros. Sans ces incidents majeurs, les assureurs ne seraient pas dans le rouge. En 2021 le ratio sinistres sur primes des grandes entreprises est revenu à la situation proche de 2019, soit 58%.

Malgré tout, si les grands groupes sont majoritairement assurés contre le risque cyber, il n'en reste pas moins que leur couverture est fortement limitée (moyenne de 38 millions d'euros) par rapport à leur exposition.

Les assureurs restent réticents tant que la mutualisation n'est pas assurée. Le réel enjeu est donc la sensibilisation des entreprises au risque cyber. Faute de mutualisation et d'historique de données disponibles très limité, les assureurs désespèrent de trouver un modèle économique viable. Les conséquences qui en découlent sont une hausse des taux de primes et des franchises et une baisse des indemnisations proposées.

En outre, les enjeux en matière d'assurance du risque cyber, exposés ci-dessus, ont motivé le choix du sujet de ce mémoire.

L'objectif de ce dernier est d'essayer d'approfondir les connaissances de ce risque en s'appuyant sur les données de la table PRC et d'étudier des modèles de prédiction de la sinistralité cyber ainsi que la dépendance entre les attaques, et ce afin d'apporter une meilleure compréhension du risque dans le cadre assurantiel.

La première partie de ce mémoire a pour objet de présenter le contexte de l'étude, le risque cyber et les problématiques posées par ce risque émergent.

La deuxième partie est consacrée à la présentation de la base des données utilisée dans le cadre de ce mémoire, suivie par l'analyse comparative des méthodes de prédiction de la gravité des sinistres issues des Séries Temporelles et des méthodes du Machine Learning.

La troisième partie est consacrée à l'étude des dépendances observées entre les sinistres et leur modélisation par les méthodes de copules.

La quatrième partie proposera une application assurantielle. Nous y aborderons une analyse de tarification de garantie contre la violation de données pour des entreprises, en se basant sur des données produites à partir de la base PRC. Cette partie mettra en application les modèles de prédiction de la gravité et du nombre de sinistres vus dans la partie 2, s'appuyant également sur des modèles issus de l'assurance paramétrique.

Partie I - Présentation générale du risque cyber

1.1 Définition du risque cyber

Il existe plusieurs définitions du risque cyber.

La norme ISO présente le risque cyber de la manière suivante : la « possibilité qu'une menace donnée exploite les vulnérabilités d'un actif ou d'un groupe d'actifs et cause ainsi un préjudice à l'organisation. Il est mesuré en termes de combinaison de la probabilité d'occurrence d'un événement et de ses conséquences ».

Le risque cyber peut aussi désigner l'ensemble des risques liés à l'usage des technologies numériques. Il peut être défini comme « un risque opérationnel portant sur la confidentialité, l'intégrité ou la disponibilité des données et systèmes d'information. Il recouvre à la fois les actes malveillants et les incidents non intentionnels issus d'erreurs humaines ou d'accidents. »[6]

Une autre définition est donnée par l'APREF[7] dans son étude sur la (ré)assurabilité du risque cyber (2016) :

« Pour toute personne morale ou physique, ci-après désignée comme l'entité.

Toutes atteintes à :

- Des systèmes électroniques et/ou informatiques [de production, d'exploitation, de gestion d'informations et de télécommunication] sous le contrôle de l'entité ou de ses prestataires ;
- Des données informatisées (personnelles, confidentielles ou d'exploitation) appartenant à ou sous le contrôle de l'entité, qu'elles soient transférées ou stockées chez elle ou chez ses prestataires.

Consécutives à :

- Un acte malveillant ou de terrorisme ;
- Une erreur humaine, une panne ou des problèmes techniques ;
- Un évènement naturel ou accidentel.

Ayant pour conséquences :

- Des dommages corporels, matériels, et/ou immatériels (frais ou pertes financières), subis par l'entité et/ou ses employés ;
- une mobilisation de ressources internes ou externes ;
- des dommages corporels, matériels, et/ou immatériels, frais ou pertes financières causés par l'entité à des tiers (y compris chaines logistiques / sous-traitants) ;
- une atteinte à la marque et/ou à la réputation de l'entité ».

De ces définitions variables découlent des stratégies variables quant au management de ce risque.

---

6 *Définition donnée dans une étude réalisée par la Direction générale du Trésor intitulée « Cyber Risk in the Financial Sector » : https://www.tresor.economie.gouv.fr/Articles/2021/12/14/cyber-risk-in-the-financial-sector*

7 *APREF : association des professionnels de la réassurance en France*

Dans le cadre de ce mémoire, on va retenir la définition donnée par Factor Analysis of Information Risk – FAIR décrivant le risque cyber comme la fréquence et l'ampleur probables d'une future perte financière résultant d'un incident cyber. Un sinistre cyber est donc tout événement affectant la confidentialité, l'intégrité ou la disponibilité du système d'information ou des données informatiques.

Les risques cyber peuvent donc résulter

- Soit d'une cyberattaque réalisée dans un but malveillant ;
- Soit d'une erreur humaine ou d'une défaillance technique.

Il existe quatre types de cyberattaques :

- L'atteinte à l'image ;
- L'espionnage : les entreprises ou même des états peuvent chercher à voler les secrets de leurs concurrents ;
- La cybercriminalité ;
- Le sabotage.

1.2 Quels sont les différents types d'attaques ?

On peut distinguer les attaques cyber par rapport à leurs objectifs, mais aussi par rapport à des approches utilisées par les pirates.

- **L'attaque par hameçonnage** ou ***phishing*** est une forme d'escroquerie sur le net, le pirate se fait passer pour un tiers de confiance (usurpation d'identité) afin d'obtenir des renseignements confidentiels pour en faire un usage criminel. Par exemple, un mail invitant à mettre à jour leurs informations personnelles sur un site falsifié. Le cybercriminel n'a plus qu'à utiliser ces informations personnelles récupérées telles que : données bancaires, mots de passe.

- **L'attaque par rançongiciel** ou **ransomware** est un logiciel malveillant qui bloque l'accès à l'ordinateur ou à des fichiers en les chiffrant et qui réclame à la victime le paiement d'une rançon pour en obtenir de nouveau l'accès. Il existe plusieurs manières d'être infecté par un ransomware : après avoir cliqué sur un lien malveillant, navigué sur des sites compromis, en ouvrant une pièce jointe infectée ou tout simplement à la suite d'une vulnérabilité logicielle.

- **L'attaque par déni de service distribué (DDoS)** est une attaque qui vise à perturber le fonctionnement normal d'un service soit par l'envoi de multiples requêtes jusqu'à la saturation du réseau, soit par une exploitation d'une faille de sécurité afin de provoquer l'interruption du réseau. L'attaque est rendue médiatique par l'attaquant pour porter atteinte à l'image et la crédibilité de la victime, en laissant penser que l'attaquant aurait pu accéder à toutes ses données, y compris les plus sensibles telles que les données personnelles, médicales, bancaires, …).

- **L'attaque Zero-day** exploite les failles de sécurité des logiciels qui n'est pas encore connue ou corrigée par l'éditeur du logiciel. Les attaquants feront tout pour multiplier le nombre de victimes tant que le correctif n'est pas déployé. La correction d'une faille de ce genre peut prendre plusieurs semaines. C'est pourquoi la mise à jour des logiciels dès que disponible est très importante pour bien se protéger de ce type d'attaque.

- **Le Malware** est un logiciel malveillant s'infiltrant dans un appareil à l'insu de son propriétaire. Il en existe plusieurs types :

    - Logiciels espionnant l'activité IT de la victime :

        - **Cheval de Troie** qui s'infiltre dans le système informatique de la victime sous l'apparence d'un logiciel légitime dans le but d'installer d'autres malware ;

        - **Keyloggers** (enregistreurs de frappe) sont des logiciels qui enregistrent tout ce qu'est tapé au clavier. Par exemple, des données d'identification aux services bancaires, ...

    - Logiciels qui ont pour objectif le chiffrement ou la suppression des données sensibles dans le but de sabotage

        - **Ransomware** chiffre les fichiers présents sur la machine et menace de tout effacer si la rançon n'est pas payée

    - Logiciels modifiants ou détournant les fonctions IT de l'appareil

        - **Le virus** peut se propager de manière incontrôlable, endommager les fonctions de base du système informatique et supprimer des données. Un virus est plus dangereux qu'un ver car il modifie les fichiers, alors qu'un vers se réplique mais n'endommage pas les fichiers.
        - **Le macrovirus** utilise les macros du Microsoft Office pour infecter un appareil
        - **Les vers** sont des programmes malveillants qui se répliquent sans cesse. Un vers se propage en occupant beaucoup d'espace et en ralentissant l'ordinateur. Les vers peuvent infecter des réseaux d'appareils, chaque machine infectée est utilisée pour en infecter d'autres.
        - **Les Rootkits** sont des logiciels qui se dissimulent dans le système infecté et qui donnent à un intrus un accès non autorisé à un ordinateur ou à un réseau. Ils sont très difficiles à détecter.

- **Le cryptojacking** est un malware forçant l'ordinateur de la victime à miner une cryptomonnaie pour le compte du pirate. Le cryptojacking peut sembler inoffensif, car la seule chose qui est volée est la puissance du processeur, mais cela pourrait ralentir et même endommager l'appareil.

1.3 Quelles sont les conséquences des incidents cyber ?

De plus en plus d'entreprises, de toutes tailles, sont touchées par des incidents cyber, qui peuvent entraîner des conséquences désastreuses et qui peuvent se répercuter sur des années. Une attaque nécessite la mise en place d'actions afin de réparer les dommages subis. Une entreprise prenant conscience des risques commencera par le rétablissement d'activité en augmentant le niveau de cybersécurité. Elle devra restaurer la confiance de ses clients tout en réglant les litiges juridiques.

Les conséquences de ces cyberattaques sont de diverses natures :

- D'ordre financier :
  - Un préjudice commercial (perte d'atouts stratégiques : perte de propriété intellectuelle)
  - Une interruption d'activité
  - Une exposition au chantage (paiement de rançon)
  - Création des failles de sécurité dans un système informatique

  Les pertes financières à la suite d'une attaque cyber peuvent être réellement conséquentes telles que la perte d'exploitation, le vol de propriété intellectuelle (brevets, …), les dépenses liées au paiement de rançon, mais aussi dépendre du montant de la réparation du matériel informatique.

- D'ordre juridique :
  - Le vol ou la perte des données (amendes réglementaires RGPD)
    Les entreprises ont l'obligation de respecter le RGPD, en cas de non-respect, elles risquent des sanctions lourdes. Cette perte peut aussi causer des frais supplémentaires de résolution des litiges tels que les frais juridiques, les honoraires d'avocat, mais aussi un préjudice à la réputation de l'entreprise.

- D'ordre réputationnel :
  - Une atteinte à l'image ou à la réputation : dépréciation de la valeur de la marque, perte de confiance accordée par le client, qui entraine une perte de revenus, une perte de marché.

Les 14 facteurs à prendre en compte pour calculer la facture d'une attaque malveillante de hackers ont bien été illustrés par l'image suivante issue de l'étude menée par le cabinet Deloitte.

Figure 1: Les 14 impacts d'une cyberattaque /Source : Deloitte

1.4 Etat réglementaire

Les entreprises sont tenues de se conformer à un certain nombre de normes en matière de cybersécurité. Le non-respect de ces normes peut donner lieu à des sanctions réglementaires.

- **Le NIST** (National Institute of Standards and Technology) est un institut national pour les standards et la technologie. C'est une division du département américain du commerce, son objectif est la promotion des normes destinées à aider les organisations dans l'évaluation des risques. FICIC[8] publiée en 2014 a pour objectif d'établir un ensemble de normes et de bonnes pratiques pour aider les organisations dans la gestion des risques liés à la cybersécurité. Le cadre de gestion des risques cyber **NIST SP 800-53** est considéré comme un standard international incontournable de la cybersécurité et surtout utilisé par des organismes fédéraux américains et des grands groupes. Cette norme de sécurité des données de 462 pages comprend un catalogue de 900 contrôles.

- **ISO/IEC 27001** est un ensemble de normes de sécurité informatique publiées par l'ISO (Organisation internationale de normalisation). Ces normes sont utilisées dans le monde

8 *FICIC : Framework for Improving Critical Infrastructure Cybersecurity*

entier. Elle annonce les exigences relatives à la sécurité des systèmes d'information. Elle donne une méthodologie permettant d'identifier, de maitriser les risques de nature cyber et de mettre en place des mesures de sécurité permettant d'assurer la disponibilité, l'intégrité et la confidentialité de l'information.

- **RGPD**[9] est aussi un instrument important au service de la cybersécurité. Ce règlement est applicable depuis le 25 mai 2018, il vise à renforcer la politique des entreprises en matière de protection des données et uniformiser le cadre légal dans le champ de la protection des données en Europe. Le RGPD est le seul texte imposant des obligations spécifiques en matière de cybersécurité et soumises au pouvoir de contrôle et de sanction d'une autorité administrative telle que la CNIL. Le RGPD établit des normes strictes sur la manière dont les organisations collectent, utilisent, gèrent, protègent et partagent les données personnelles. Les amendes pour violation du règlement sont très élevées et peuvent atteindre 20 millions d'euros ou 4 % du revenu annuel global d'un groupe (le montant le plus élevé est retenu). L'une des obligations les plus importantes du RGPD est l'obligation de notifier aux autorités de contrôle dans les 72 heures toute violation de données susceptible de mettre en danger les données personnelles.

- **PCI DSS**[10] est une norme de sécurité des données obligatoire pour la plupart des entreprises qui collectent, traitent et stockent les données des cartes de paiement.

- **La Directive NIS** est une directive européenne sur la sécurité des réseaux et des systèmes d'information. Cette directive s'applique aux entreprises des secteurs stratégiques comme les transports, énergie, la finance, la santé, l'eau, les infrastructures essentielles. Elle met en place une obligation de notifier les incidents à l'autorité nationale compétente. La directive NIS 2 devra couvrir un plus grand nombre de secteurs. Son objectif est à renforcer les obligations des entreprises en matière de sécurité informatique grâce à des mesures de contrôle plus strictes de la part des autorités nationales.

- **Le Cybersecurity Act** est entré en vigueur le 27/06/2019 crée un cadre légal et un cadre européen harmonisé de certification de cybersécurité.

- Le règlement **DORA** « **Digital Operational Resilience Act**» est entré en vigueur le 16/01/2023. Il s'appliquera à tous les membres de l'Union Européenne à partir du 17/01/2025 et prévoit de passer d'une approche préventive à une approche pro-active dans la gestion des risques opérationnels. Ce règlement européen vise à rendre le secteur bancaire et financier plus résistant au piratage. Son objectif est d'établir des règles communes en matière de sécurité informatique : gestion des risques informatiques, déclaration des incidents cybers majeurs notamment.

1.5 Attaques célèbres

Quelles que soient les motivations derrière les attaques financières ou politiques, les cyberattaques peuvent entraîner des conséquences désastreuses. Au 21ème siècle, la

9 *RGPD : Règlement Général sur la Protection des Données*
10 *PCI DSS : Payment Card Industry Data Security Standard*

cybersécurité est devenue une considération géopolitique de plus en plus vitale. En cas de violation, les résultats peuvent être catastrophiques.

- **SolarWinds cyberattack (2020)** une infiltration informatique d'une ampleur inédite. En septembre 2019, les pirates qui seraient dirigés par une opération d'espionnage russe, ont infiltré le système en insérant un code malveillant dans une mise à jour d'Orion, le logiciel phare de SolarWinds, ce logiciel est utilisé par des milliers d'entreprises et d'organisations à travers le monde. De cette manière, ils ont obtenu un accès illimité à des milliers d'organisations et le gouvernement américain pendant plus d'une année. Des grandes entreprises telles que Microsoft, les départements du gouvernement américain stratégiques dont ceux de l'Energie, du Commerce, du Trésor ont été touchés.
- **Not Petya (2017)** est un malware qui a commencé à agir en Ukraine avant de se propager partout dans le monde. Il a contaminé des milliers d'ordinateurs, perturbant pendant plusieurs mois des multinationales et infrastructures critiques. Selon l'ancien conseiller de la sécurité intérieure de la Maison Blanche, Tom Bossert, le coût de cette attaque Not Petya était supérieur à 10 milliards de dollars. La finalité de cette attaque n'était pas l'extorsion, mais plutôt la destruction des données. Le mobile derrière cette attaque était plutôt politique, que financier et l'Ukraine en était la cible principale.
- **WannaCry (2017)** est une attaque par ransomware, une faille du système Microsoft Windows qui avait été exploitée. Faute d'avoir mis à jour Windows avant l'attaque, les victimes n'ont pas pu bénéficier du correctif et se sont retrouvés bloqués hors de leur système. Le paiement d'une rançon en Bitcoin leur a été exigé pour en retrouver l'accès. Cependant le code utilisé était défectueux et donc même en payant la rançon les victimes n'ont pas pu récupérer leurs fichiers, puisque les attaquants n'avaient aucun moyen pour associer le paiement de la rançon à l'ordinateur de la victime.
- **RockYou2021 (2021)** : divulgation sur un forum de hackers de la plus grande collection (dictionnaire) de mots de passe de tous les temps. Il s'est avéré que la grande majorité des 8,4 milliards de mots de passe prétendument divulgués étaient déjà bien connus - la liste était une énorme compilation et n'a révélé aucun mot de passe fraîchement compromis.
- **Colonian Pipeline (2021)** : Colonial Pipeline est l'un des plus grands opérateurs de gazoducs américains, il a dû cesser toutes ses opérations en raison d'une cyberattaque, qui a eu pour conséquence l'arrêt de distribution de 45% du carburant sur la côte Est des États-Unis.

1.6 Les caractéristiques de l'assurabilité du risque cyber

Le marché spécialisé en assurance cyber souffre. Les compagnies d'assurance perdent de l'argent et les entreprises ont du mal à s'assurer. Le risque cyber présente plusieurs problèmes d'assurabilité. Un sondage réalisé par l'AMRAE auprès des entreprises montre que les

entreprises rencontrent des difficultés pour souscrire et même renouveler leurs couvertures cyber. Les assureurs exigent des hausses tarifaires, augmentation des franchises, réduction des risques assurés et voire refus de reconduction des contrats. Les assureurs font face au problème de capacité. La capacité est un engagement financier maximal de l'assureur. La branche cyber n'est pas rentable, les assureurs ont remboursé plus qu'ils en ont encaissé. Quelles seraient les causes de cette frilosité de la part des assureurs ?

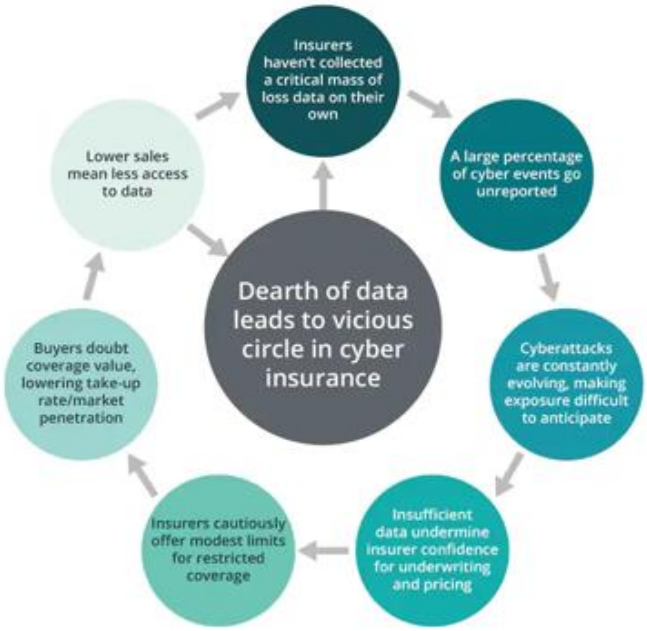


Figure 2 : Le cercle vicieux du risque cyber / Source : Deloitte

- **Historique de données limité**

  Le risque cyber est un risque récent. Le manque de données ne permettant pas de bien quantifier les risques cyber. Modéliser un tarif devient un défi pour les assureurs. Les entreprises préfèrent cacher leurs incidents cyber pour pouvoir préserver leur réputation. Il y a donc peu de recul sur la fréquence et la sévérité des sinistres de type cyber. En France, il n'existe pas pour le moment de base de données fiable. Seule l'obligation de notifier les vols des données dépassant un certain seuil permet d'avoir un peu d'historique.

- **Mutualisation faible**

L'entrée sur le marché de la plus grande masse des entreprises est donc nécessaire pour pouvoir mutualiser ce risque. Les entreprises qui ont conscience de ce risque ne sont pas très nombreuses. Certaines entreprises n'ont même entamé aucune démarche sur la cybersécurité. Comme dit précédemment, 87% des grandes entreprises étaient couvertes en 2020, mais seulement 8% des ETI, 1% des communes de plus de 5000 habitants et 0,0026% des PME. Le montant cumulé des primes ne représentait que 0.225% de l'ensemble des primes en assurance non-vie.
Ce sont les grandes entreprises qui ont eu des sinistres qui ont coûté le plus aux assureurs. Cela entraine un manque d'offre du côté des grandes entreprises et une faible demande de la part des PME et des ETI.
On assiste donc à un cercle vicieux : l'offre n'est plus attractive, donc il y a moins de clients, donc moins de mutualisation, donc moins de capacité.

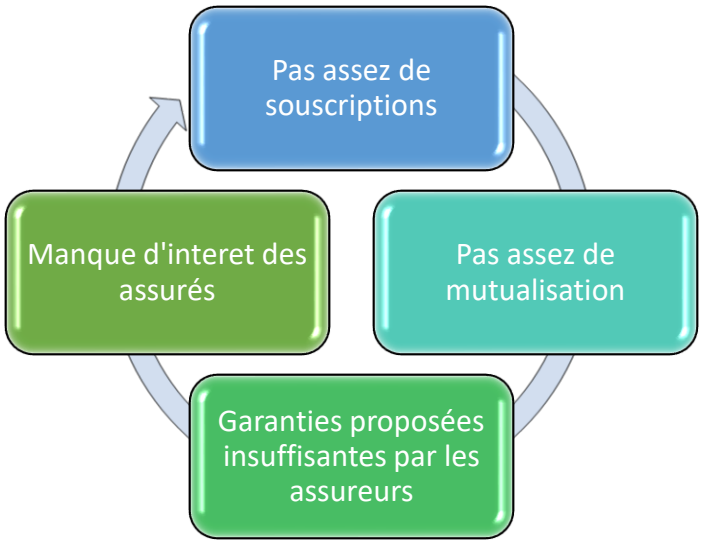


Figure 3 : Cercle vicieux de l'assurance cyber bis

- **Risque fortement évolutif**

Le risque cyber se trouve en perpétuelle évolution et est donc difficile à appréhender. Les pirates ne cessent d'adapter leurs modes opératoires. Les hackers, pour parvenir à leurs fins, n'exploitent pas seulement des failles de sécurité, mais aussi des relations de confiance entre les parties prenantes. Aucun système de sécurité ne peut garantir le risque zéro.

- **Problème d'antisélection**

  Le risque cyber présente un problème d'antisélection, car l'assureur dispose habituellement de peu d'informations sur le niveau de sécurité adopté par l'entreprise. Sans faire appel aux experts en cybersécurité, Il n'a pas de moyens de distinguer les bons risques des mauvais risques. Il y a des critères liés à l'entreprise tels que sa localisation, secteur d'activité, taille, nombre d'ordinateurs connectés, architecture du réseau informatique et bien d'autres qui peuvent aider les assureurs à mettre en place un scoring permettant une meilleure mesure de l'exposition au risque cyber. Mais même en ayant le niveau de sécurité le plus élevé possible, le risque zéro n'est pas atteignable, puisqu'on doit toujours prendre en compte le facteur humain.

- **Risque quasi systémique**

  Un risque systémique est un risque qui peut se propager au sein d'un système, mais aussi d'un système à l'autre à travers des interconnexions. Il n'existe aucune frontière géographique pour la menace cyber. Les pirates peuvent attaquer les ordinateurs se trouvant à des milliers de kilomètres. L'effet des attaques est accentué par l'homogénéité des systèmes informatiques et l'interdépendance des niveaux de sécurité. La viabilité des assureurs pourrait être menacée par une catastrophe cyber touchant un nombre considérable d'acteurs économiques. Ce sont les risques extrêmes qui ont cet aspect systémique, qui n'est pas encore bien appréhendé par les assureurs (y compris les réassureurs), qui ne disposent probablement pas de capacité pour y faire face. On ne peut pas estimer le montant à indemniser en cas de scénario catastrophe. Le chiffre pourrait être colossal. Le risque cyber n'est donc pas totalement assurable en l'état actuel des connaissances.

Partie II – Prédiction de la sinistralité du risque Cyber – analyse comparative des méthodes de Séries Temporelles vs. Méthodes de Machine Learning

Cette partie a pour objectif de présenter une analyse comparative des méthodes les plus utilisées des séries temporelles et du Machine Learning.

La première section présente la base de données américaine PRC suivie des statistiques descriptives (section 2.1), de la procédure de prétraitement effectué (section 2.2).

La troisième et la quatrième section présenteront des résultats de la modélisation de la sévérité d'un incident cyber par les séries temporelles (section 2.4) et par les méthodes du Machine Learning (section 2.3).

2.1 Description de la base des données PRC

L'analyse est basée sur l'ensemble de données fournie par le Privacy Right Clearinghouse (PRC), qui est l'une des plus grandes bases de données disponibles publiquement depuis 2005 et qui a été largement étudiée dans la littérature.

Elle ne répertorie que les fuites des données affectant des citoyens américains. La majorité des sinistres est donc située aux Etats Unis.

Le RGPD définit une violation de données comme « une violation de la sécurité entraînante, de manière accidentelle ou illicite, la destruction, la perte, l'altération, la divulgation non autorisée de données à caractère personnel transmises, conservées ou traitées d'une autre manière, ou l'accès non autorisé à de telles données. » La définition donnée par la CNIL est la suivante : « tout incident de sécurité, d'origine malveillante ou non et se produisant de manière intentionnelle ou non, ayant comme conséquence de compromettre l'intégrité, la confidentialité ou la disponibilité » de données personnelles.

Cette base contient 8927 enregistrements, le dernier en date est celui du 25/10/2019. Les fuites de données recensées ont été rapportées par des agences gouvernementales ou via des sources médiatiques. La base PRC comporte 11 variables, dont 2 variables spatiales (Longitude, Latitude), une variable indiquant la date de notification de l'incident et 8 qualitatives, excepté la variable cible indiquant le nombre d'enregistrements compromis par une fuite :

Nom variable	Description
Date made Public	Date à laquelle une fuite a été déclarée publiquement
Company	Nom de l'entreprise touchée par la fuite
City	Ville où l'entreprise est localisée
State	Etat où l'entreprise est localisée
Type of Breach	Type de violation de données (8 modalités)
HACK	*Piraté par un tiers ou infecté par un logiciel malveillant*
CARD	*Fraude impliquant des cartes de débit et de crédit pas accompli via le piratage*
INSD	*Interne : Divulgation intentionnelle d'informations sensibles par un individu avec droit d'accès légitime*
PHYS	*Documents papier perdus, jetés ou volés (non électroniques)*
PORT	*Matériel électronique volé, jeté ou perdu : ordinateur portable, PDA, smartphone, clé USB, CD, disque dur, …*
STAT	*Perte ou vol d'ordinateur fixe*
DISC	*Divulgation involontaire (n'impliquant pas de piratage, de violation intentionnelle ou de perte physique)*
UNKN	*Type de fuite inconnu*
Type of Organisation	Secteur d'activité de l'entreprise qui a subi la fuite (8 modalités)
BSF	*Entreprises du secteur financier ou assurances*
BSO	*Entreprises d'autres secteurs*
BSR	*Secteur commercial (commerces de détail en ligne y compris)*
EDU	*Etablissements d'enseignement*
GOV	*Gouvernement et armée*
MED	*Etablissements du secteur médical*
NGO	*Organismes à but non lucratif*
UNKN	*Secteur d'organisation inconnu*
Total Records	Nombre de données violées (nombre d'enregistrements)
Description of Incident	Description de l'incident qui s'est produit
Information Source	Source dont provient l'information
Year of Breach	Année de survenance de l'incident
Longitude	Position géographique
Latitude	

Figure 4 : Description des variables de la base PRC

La base PRC ne dispose pas de coût financier résultant d'un incident cyber. Cependant, le nombre d'enregistrements est un élément clé pour pouvoir mesurer la sévérité de l'incident. La variable cible (Total Records) est très volatile, ses valeurs se répartissent entre 0 et 3 milliards. La moyenne du nombre d'enregistrements fuités par un incident est très supérieure à la médiane, c'est dû à un nombre élevé des valeurs extrêmes. Sur le graphe en dessous, on peut observer le décalage entre les valeurs de la moyenne et la médiane.

	Moyenne	Médiane
Tous les sinistres (y compris ceux dont le Nb of records =0)	1 218 677	946
Sinistres dont le Nb of records > 0	1 644 456	2029

Figure 5 : Tableau comparatif - moyenne, médiane d'un incident cyber

La figure suivante représente la répartition des incidents par type de fuite : les types de fuite les plus représentées sont HACK (30%), DISC (22%), PHYS (21%) et PORT (14%), c'est-à-dire les fuites les plus courantes dans cette base sont le piratage (HACK), la divulgation involontaire (DISC), la perte ou le vol de documents papier (PHYS) et le matériel électronique volé ou perdu (PORT).

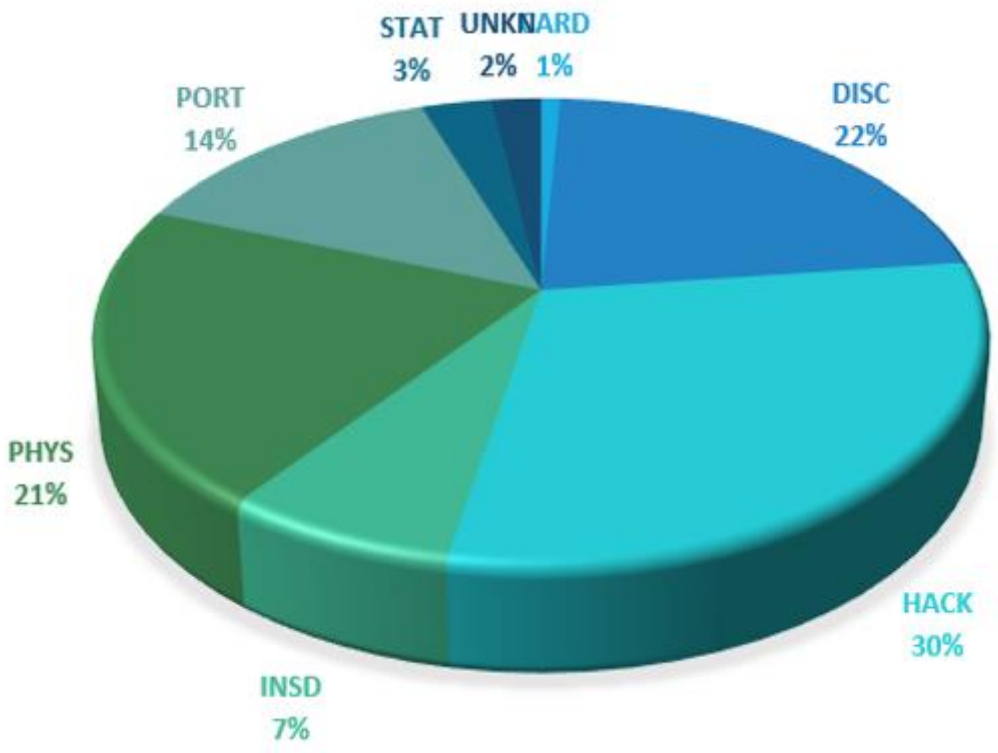


Figure 6 : Répartition des incidents cyber par type de fuite

Les Figures 5 et 6 représentent la répartition des incidents par secteur d'activité : elle met en lumière que plus de moitié des organisations attaquées fait partie du secteur médical (52%). Ce n'est pas étonnant, car les données médicales sont très recherchées sur le darknet. Une autre explication possible est l'existence aux Etats-Unis d'une obligation légale de notifier les fuites des données affectant plus de 500 personnes dans le secteur médical. La base PRC présente

donc un biais, cela rend difficile l'estimation de la fréquence essentielle pour estimer un tarif. Ce biais a été mis en évidence dans les travaux de Sébastien Farkas, Olivier Lopez et Maud Thomas[11]. Il y a également une forte hausse des sinistres depuis 2010, ils ont quasiment doublé, comme on peut le voir sur la figure 3. Il reste difficile à déterminer la proportion de sinistres qui est causée par l'évolution de la menace et celle provoquée par la modification du processus de notification des sinistres dans certains domaines imposée par le gouvernement américain.

Les autres secteurs d'activité se répartissent uniformément autour de 10%, excepté les organisations à but non lucratif.

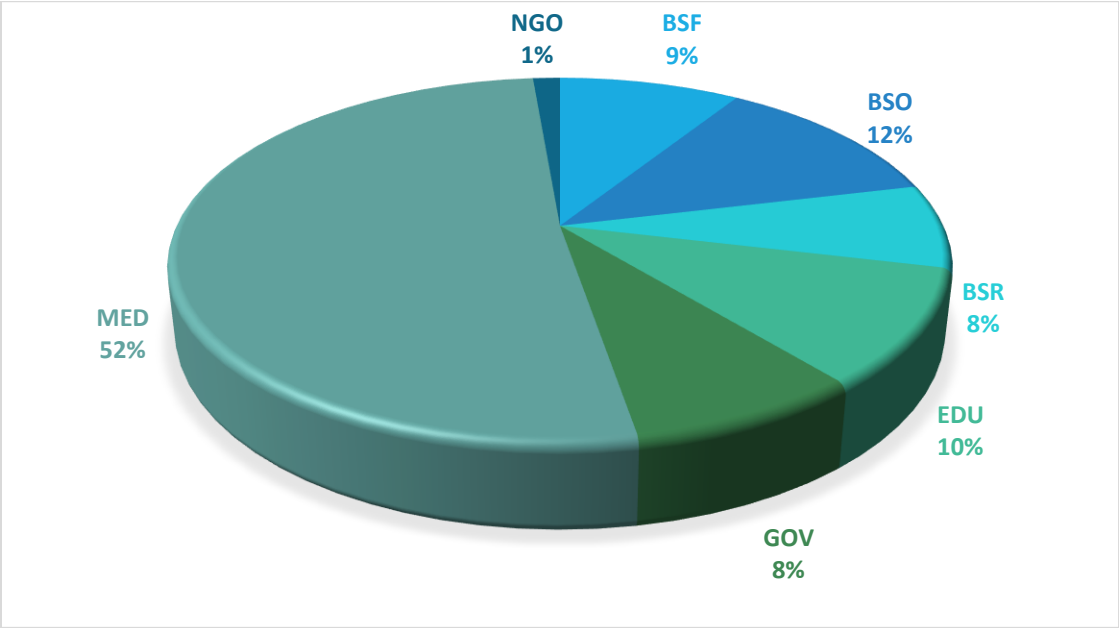


Figure 7 : Répartition des incidents par secteur d'activité

Les Etats américains les plus touchés par le problème des fuites des données sont : la Californie (1325), New York (615), le Texas (577) et la Floride avec 451 incidents répertoriés. On peut voir cette répartition dans le tableau en annexe et le graphe en dessous.

11 *Cyber claim analysis through Generalized Pareto Regression Trees with applications to insurance Sébastien Farkas, Olivier Lopez, Maud Thomas*

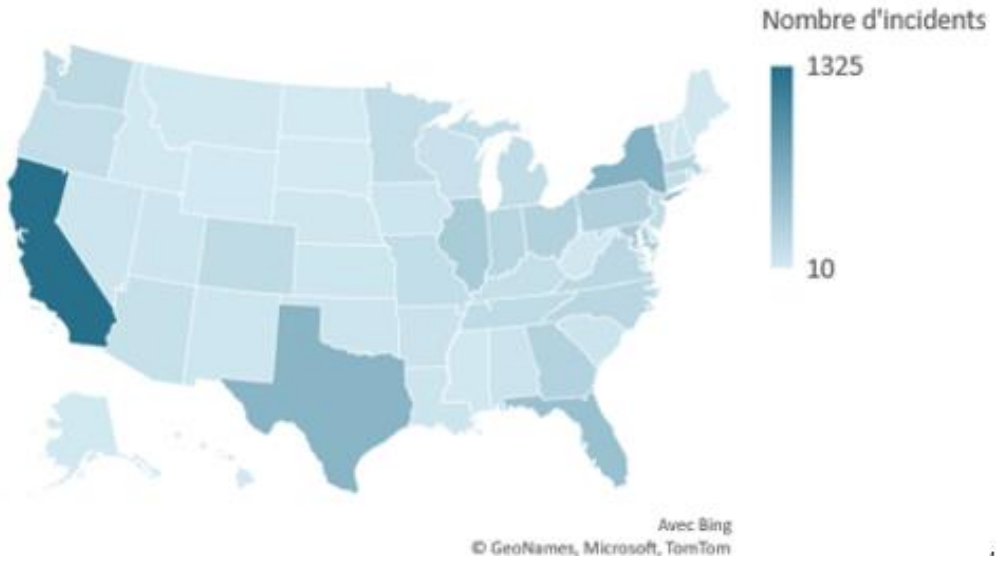


Figure 8 : Nombre total d'incidents cyber par Etat

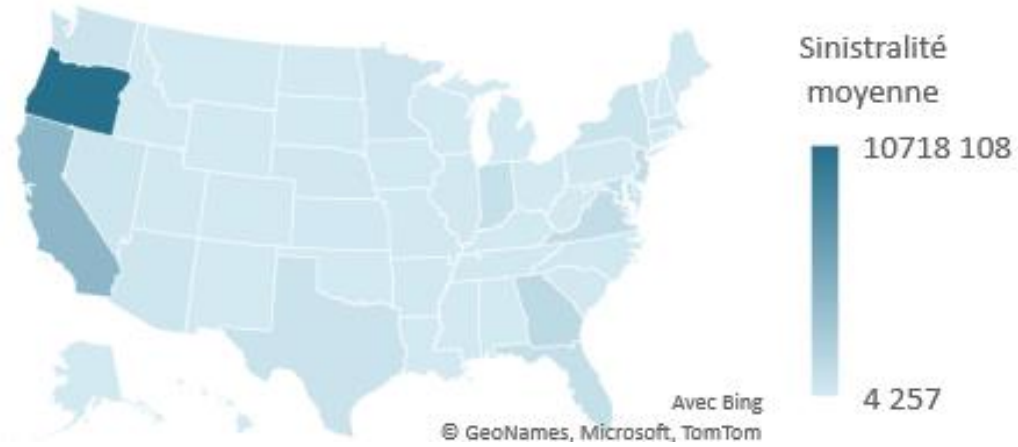


Figure 9 : Cartographie de la sinistralité moyenne par Etat (2005-2018)

Dans le but de réaliser une répartition plus détaillée de la sévérité extrême des incidents cyber, le recodage suivant de la variable a été effectué :

```
Breaks = c (-Inf., 100 000, 500 000, 1 000 000, 15 000 000, 500 000 000, Inf.)
Labels = c ("1","2", "3","4","5","6")
```

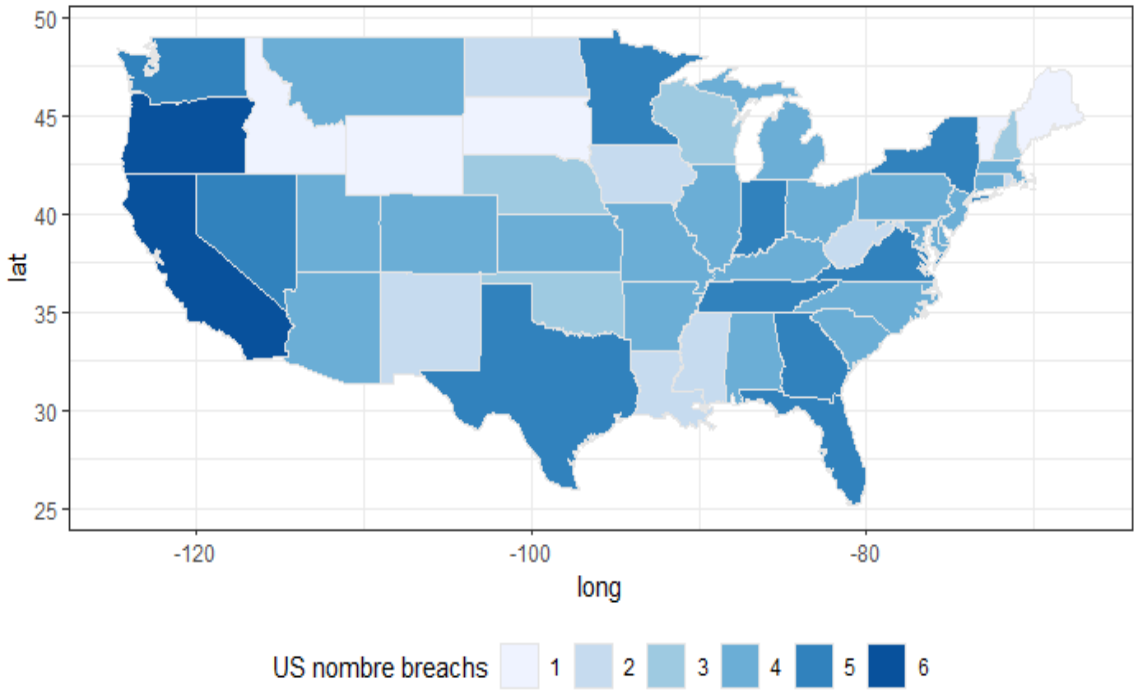


Figure 10 : Répartition de la sévérité totale recodée en classes entre 2005 et 2018

Les états américains touchés par les incidents les plus graves et coûteux sont les états de la côte Ouest : Californie, Oregon, Washington, mais aussi ceux de la cote Est : Texas, New-York, Floride.

2.2 Prétraitement de la base des données PRC

Dans cette base de données, une valeur manquante ou nulle ne serait due qu'à des problèmes de déclaration, c'est-à-dire une information indisponible, plutôt qu'à une fuite de 0 enregistrement. Dans le cadre de cette étude, seuls les incidents avec une valeur positive du nombre d'enregistrements compromis sont sélectionnés car la plupart des modèles de gravité utilisés dans le secteur de l'assurance supposent une valeur positive du sinistre. Par conséquent, toutes les observations manquantes ou nulles sont supprimées. De plus, seuls les incidents localisés dans la zone continentale des États-Unis sont pris en compte.

Sur les graphes en dessous, on peut observer une augmentation significative du nombre des incidents et la prévalence du secteur médical à partir de 2010 (doublement), on peut supposer que cela est dû à l'augmentation des incidents, mais aussi à une obligation de notification à partir d'un certain niveau de gravité.

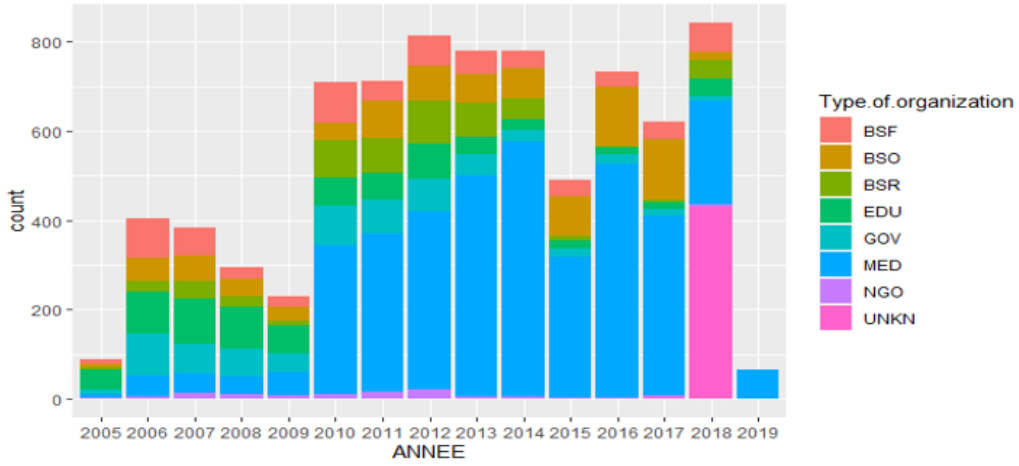


Figure 11 : Répartition par type d'organisation touchée par année

Par ailleurs, en 2018, près de 50% des incidents ont été saisis comme provenant d'une violation de type inconnu (UNKN), il s'agît d'une anomalie dans l'alimentation de la base.

Pour toutes ces raisons, il a été décidé d'éliminer les années atypiques du cadre de l'étude (2005 à 2009 et 2018 et au-delà) :

- 2005 à 2009 : observations très différentes à partir de 2010. Période non représentative de la situation la plus récente. Elle pourrait apporter un biais dans les analyses (fréquence et sévérité des évènements sensiblement différentes avant et à partir de 2010)
- 2018 : année atypique dont les observations remontées dans la base PRC apporteraient un biais dans l'analyse.
- 2019 : année incomplète et également atypique (prévalence du secteur médical), qui apporterait un biais dans l'analyse. La période sélectionnée pour cette étude est donc celle allant du janvier 2010 à décembre 2017.

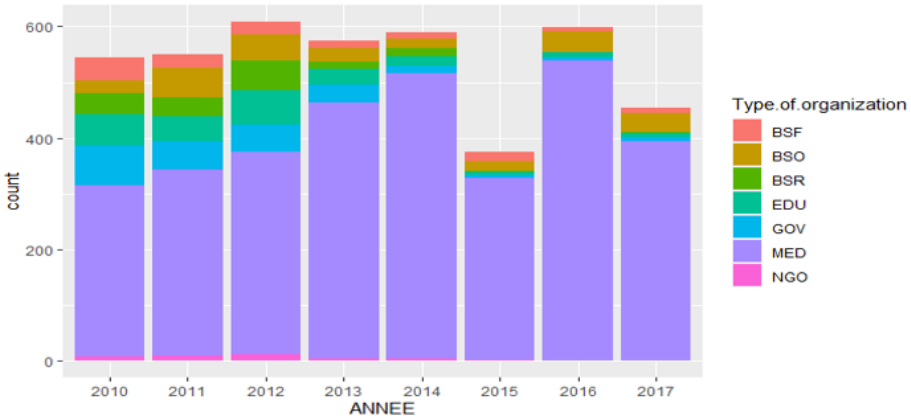


Figure 12 : Répartition par type d'organisation touchée par année (2010 à 2017)

On peut se demander si la période sélectionnée est réellement pertinente, puisque le risque cyber ne cesse d'évoluer depuis quelques années. La base PRC reste limitée, il est évident qu'elle ne reflète pas la situation réelle concernant le nombre de fuites observées, il y a beaucoup plus d'incidents réellement survenus que d'incidents déclarés. Les entreprises sont réticentes à déclarer des fuites des données pour éviter une mauvaise publicité. De plus, la date de déclaration ne correspond pas forcément à la date à laquelle l'incident cyber s'était produit. Cependant elle reste la base publique disponible la plus complète à ce jour. Le manque de disponibilité de l'historique reste un problème majeur pour les chercheurs travaillant sur cette thématique.

2.3 Mise en place des méthodes du Machine Learning dans la prédiction de la sévérité des incidents cyber

2.3.1 Méthodes de classification

Pour pouvoir utiliser la majorité des méthodes du Machine Learning, la variable cible représentant le nombre d'enregistrements compromis par un incident cyber a été recodée en 4 modalités de la manière suivante :

Modalité	Range
XS	[1,500]
S	[501,10000]
M	[10001,100000]
L	>100000

Figure 13 : Modalités de recodage de la variable Nb of Records

Cette section se concentre sur les algorithmes d'apprentissage automatique ML utilisés dans l'objectif de trouver l'algorithme le plus performant pour la prédiction de la gravité d'un sinistre cyber. Chaque modèle est brièvement présenté pour permettre au lecteur de comprendre son mode de fonctionnement. Ensuite les méthodes seront comparées afin de sélectionner un modèle le plus performant au sens des critères de validation choisis.

Les algorithmes de Machine Learning peuvent être regroupés en deux grandes catégories : l'**apprentissage supervisé** et l'**apprentissage non supervisé**. Le choix dépend de la typologie de l'information à traiter. A la différence de l'apprentissage supervisé, dans l'apprentissage non supervisé, nous ne disposons pas de données Yi (variable à expliquer).

Les modèles du ML testés sont :

- Les Arbres de décision,
- Le Random Forest,
- Le Naïve Bayésien,
- Le GLM,
- Les Réseaux de neurones
- Et le XGBOOST.

2.3.1.1 Apprentissage supervisé

Pour éviter le surapprentissage, le jeu de données est séparé en deux échantillons distincts : un ensemble d'apprentissage (train) et un ensemble de test. L'échantillon de test est obtenu par une sélection aléatoire des lignes de la base de données totale.

Les algorithmes appliquent un modèle à un ensemble de données d'apprentissage (train) afin de prédire :

- Une classe (dans la classification) ;
- Une valeur numérique (dans la régression).

Le but de cet apprentissage est de s'entrainer sur un ensemble de données et de pouvoir l'appliquer à des nouvelles données. Les échantillons utilisés pour l'apprentissage sont bien différents de ceux utilisés pour tester. De cette manière, on pourrait mesurer la capacité du modèle à généraliser ce qu'il a appris. On peut détecter le surapprentissage en voyant un modèle ayant des performances du jeu d'apprentissage largement supérieures à celles de l'échantillon du test.

Dans cette section, les algorithmes suivants seront testés :

- Les Arbres de décision ;
- Le Random Forest ;
- Le Gradient boosting (XGBOOST) ;
- Le GLM ;
- Le Naïve Bayésien ;
- Les Réseaux de neurones.

2.3.1.2 Arbres de décision

Les arbres de décision est une méthode d'apprentissage supervisé non paramétrique et qui peuvent être utilisés à la fois pour des tâches de classification et de régression. Comme son nom l'indique il possède une structure à la fois hiérarchique et arborescente : il se compose d'un nœud racine, de branches, de nœuds internes et de nœuds feuilles. Dans cette représentation, chaque nœud interne correspond à un attribut, et chaque feuille correspond à une classe.
C'est un algorithme qui segmente l'ensemble de données par des coupes successives. À chaque nœud, l'algorithme détermine la meilleure variable à utiliser et à quel point couper afin d'obtenir le meilleur critère qui séparera la population en deux sous-ensembles les plus homogènes possibles. L'algorithme de CART peut être représenté par la figure suivante.

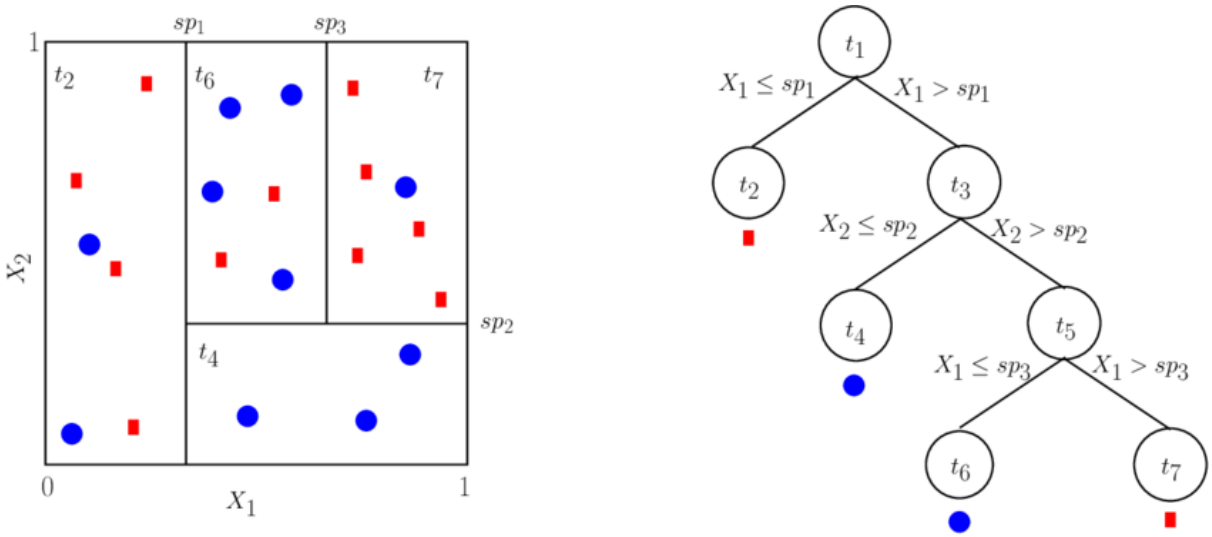


Figure 14 : Représentation de l'algorithme CART

Un arbre de décision est un algorithme itératif qui divise les individus en k groupes à chaque itération pour pouvoir expliquer la variable cible. Si k=2 il s'agit d'un arbre binaire.

L'algorithme de construction d'un arbre de décision permet :

- De choisir une variable explicative dans l'ensemble d'apprentissage, qui par ses modalités, permet la meilleure séparation des individus en sous-ensembles (nœuds), en maximisant la variance inter-groupe.
- La réitération de ce même processus de séparation pour chaque sous-population jusqu'à l'arrêt de ce processus (critère d'arrêt à définir). C'est-à-dire on va générer de nouveaux arbres de décision récursivement en utilisant les sous-ensembles créés. On s'arrête lorsqu'on ne peut plus classifier, on obtient les nœuds finaux : les feuilles.

La moyenne des valeurs de chaque individu de la feuille nous donne le coût de chaque feuille.

L'algorithme CART (arbre de classification et de régression) a été développé en 1984 par L. Breiman, J. Friedman, C. Stone et R. Olshen. L'algorithme CART est conçu pour construire un modèle binaire de l'arbre de décision. Il supporte la régression.

La meilleure caractéristique est déterminée en utilisant la fonction d'entropie et le gain d'information.

Soit un ensemble de classes C, l'entropie de l'ensemble de données S est donnée par la formule suivante :

$$H(S) = \sum_{c_i \in C} -P(c_i)\, log_2 P(c_i)$$

Où :

$$P(c_i) = \frac{|c_i|}{|S|}$$

Soit f est un vecteur des variables et f_j les valeurs de cette variable. On peut donc séparer les données S en plusieurs sous-ensembles S_j Le gain d'information est estimé sur la base de la différence entre l'entropie d'origine de S et celle mesurée après la division en fonction de la caractéristique f_j .

$$IG(S, f_j) = H(S) - \sum_{S_{jk} \in S_j} P(S_{jk}) H(S_{jk})$$

Où :

$$P(S_{jk}) = \frac{|S_{jk}|}{|S|}$$

La variable sélectionnée est celle ayant plus de gain d'information. La valeur avec entropie nulle est considérée être une feuille de l'arbre.

```
Confusion Matrix and Statistics

          Reference
Prediction  XS   S   M   L
        XS  95  58  20  17
        S   37 463 102  35
        M    2   5   7   6
        L    3   3   3   3

Overall Statistics

               Accuracy : 0.6612
                 95% CI : (0.6285, 0.6929)
    No Information Rate : 0.6158
    P-Value [Acc > NIR] : 0.003266

                  Kappa : 0.3271

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: XS Class: S Class: M Class: L
Precision               0.5000   0.7268 0.350000 0.250000
Recall                  0.6934   0.8752 0.053030 0.049180
F1                      0.5810   0.7942 0.092105 0.082192
Prevalence              0.1595   0.6158 0.153667 0.071013
Detection Rate          0.1106   0.5390 0.008149 0.003492
Detection Prevalence    0.2212   0.7416 0.023283 0.013970
Balanced Accuracy       0.7809   0.6740 0.517574 0.518951
```

L'accuracy permet de décrire la performance du modèle. Elle mesure le taux de prédictions correctes sur l'ensemble des individus, c'est-à-dire la prédiction des individus positifs et négatifs. Dans notre cas, elle est égale à 66.12%, donc le modèle prédit de manière exacte dans 66.12% des cas. On peut observer que ce sont les classe M et L, qui ont le taux de prédiction le plus bas. Les modalités de la variable représentant la gravité du sinistre sont déséquilibrées.

La classe S est surreprésentée en nombre.

L'utilisation de l'accuracy n'est donc pas la plus adaptée pour ce type de données.

Modalité	Range	Nombre
XS	[1,500]	662
S	[501,10000]	2632
M	[10001,100000]	728
L	>100000	270

Les métriques telles que la précision, le recall ou la courbe ROC AUC sont plus adaptées.

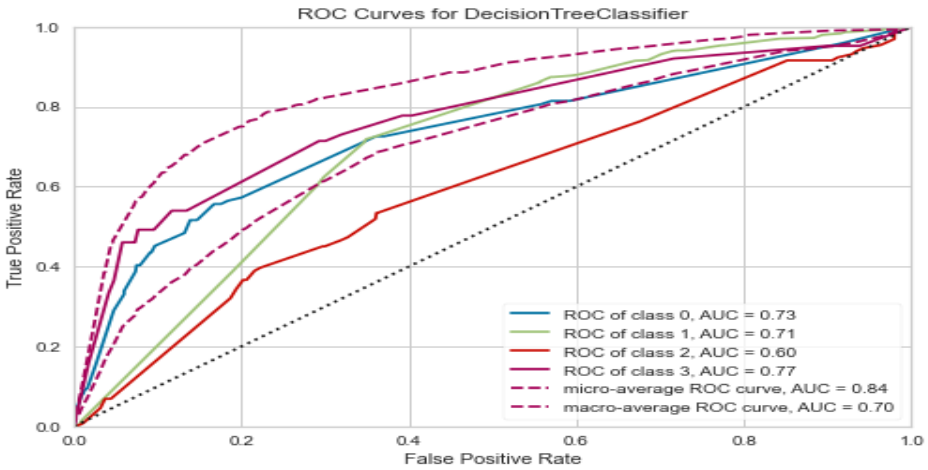


Figure 15 : Courbe ROC AUC (arbres de décision)

La solution possible pour le problème des données déséquilibrées sont des techniques de sous ou sur-échantillonnage. Par exemple, SMOTE (Synthetic Minority Oversampling Technique) est une méthode de suréchantillonnage des observations minoritaires pour leur donner plus d'importance lors de la modélisation.

Parmi les avantages des arbres de décision, on peut citer :

- Méthode rapide en temps de calcul et facile dans l'interprétation des résultats.
- Elle peut gérer tous types de variables et permet aussi de choisir les plus pertinentes.

Cependant, cette méthode présente aussi quelques inconvénients. Le pouvoir prédictif du modèle dépend fortement de l'ordre dans lequel les variables sont choisies. Afin de remédier à ce problème, Les techniques telles que le boosting ou le bagging peuvent être utilisées.

2.3.1.3 Random Forest (Foret aléatoire)

L'algorithme du Random Forest a été proposé par Léo Breiman en 2001. Il est basé sur l'assemblage d'arbres de décision, il s'agit du cas particulier de bagging. L'idée de cet algorithme est assez simple : plutôt que d'utiliser un estimateur complexe, l'algorithme du Random Forest se base sur plusieurs estimateurs simples, mais de précision plus faible. Si un estimateur présente un point de vue du problème, l'assemblage de tous ces estimateurs nous donne une vision globale. On construit donc un ensemble d'arbres de décision sur des échantillons Bootstrap, puis des prédictions obtenues sont agrégées. Pour chaque scission, plutôt que de chercher la meilleure coupe parmi toutes les variables explicatives (n), on sélectionne la meilleure coupe pour p variables explicatives tirées aléatoirement parmi n.

Une formule simple qui permet de bien illustrer le Random Forest est :

Forest = tree bagging + feature sampling

- Le bagging signifie « bootstrap aggregation » : en réalisant un échantillonnage des données il entraîne l'algorithme sur chacun de ces n échantillons séparément. Il

assemble ensuite les résultats des modèles obtenus en prenant la majorité parmi les n prévisions.

- Feature sampling : on tire aléatoirement sur les variables. Par défaut, on utilise $\sqrt{n}$ variables lorsque le nombre total des variables est n en cas de classification et n/3 en cas de régression. Ce procédé permet de diminuer la corrélation entre les variables explicatives et donc réduit la variance.

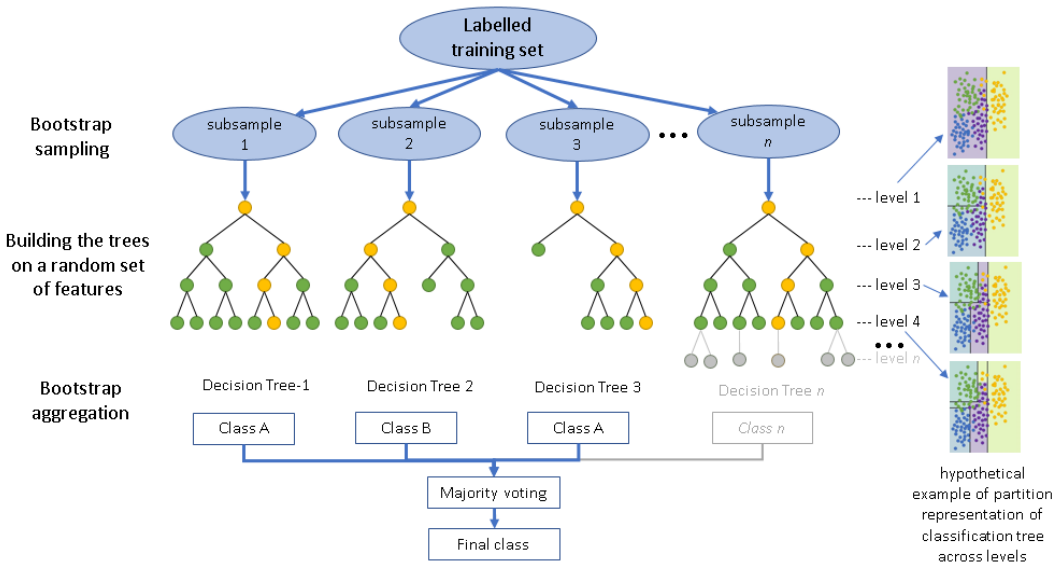


Figure 16 : L'algorithme de random forest / Source : https://medium.com/

L'accuracy est égale à 66.9%, elle est en peu plus élevé que celles des arbres de décision. Il reste toujours le problème de données déséquilibrées. On va donc aussi voir la courbe ROC.

```
Confusion Matrix and Statistics

          Reference
Prediction  XS   S   M   L
        XS  66  28  16   6
        S   56 477 107  30
        M    9  18  18   5
        L    1   3   5  13

Overall Statistics

               Accuracy : 0.669
                 95% CI : (0.6364, 0.7004)
    No Information Rate : 0.6131
    P-Value [Acc > NIR] : 0.0003894

                  Kappa : 0.323

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: XS Class: S Class: M Class: L
Precision              0.56897   0.7119  0.36000  0.59091
Recall                 0.50000   0.9068  0.12329  0.24074
F1                     0.53226   0.7977  0.18367  0.34211
Prevalence             0.15385   0.6131  0.17016  0.06294
Detection Rate         0.07692   0.5559  0.02098  0.01515
Detection Prevalence   0.13520   0.7809  0.05828  0.02564
Balanced Accuracy      0.71556   0.6628  0.53917  0.61477
```

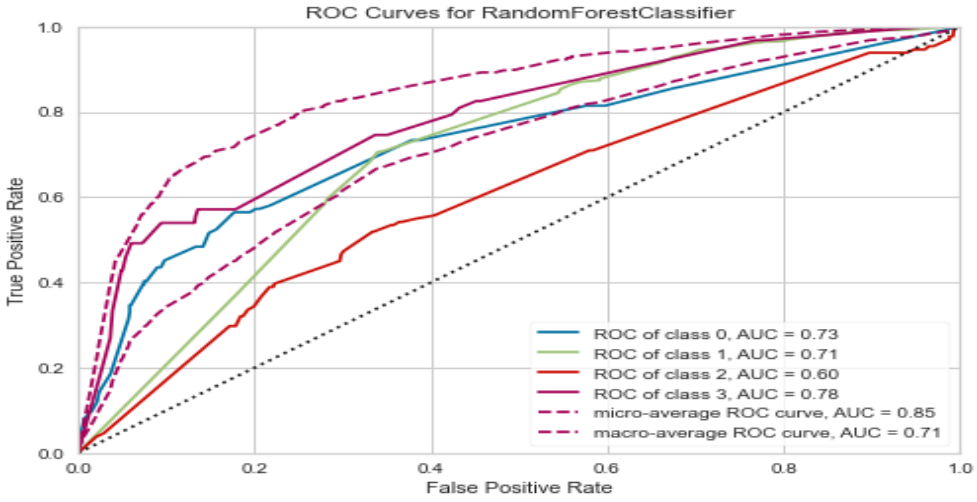


Figure 17 : Résultats et Courbe ROC AUC (Random Forest)

Parmi les avantages, on retrouve :

- Il n'est pas enclin au surapprentissage.
- Il est très intuitif, facile d'utilisation.
- Il n'y a pas de nécessité de « mettre à l'échelle » les données.
- Il est efficace même sur les grandes bases de données.

Ses inconvénients sont les suivants :

- Il est plus « gourmand » en temps de calcul que l'algorithme d'arbre de décision.
- C'est une boite noire, donc l'explicabilité du modèle est plus faible.
- Le résultat du Random Forest peut changer considérablement par un petit changement dans les données.

2.3.1.4 Gradient Boosting

Actuellement, le gradient boosting est l'une des techniques les plus puissantes pour construire des modèles prédictifs. C'est une méthode permettant de réduire les erreurs lors de la prédiction, en augmentant le poids des observations qu'on trouve difficiles à classer et en diminuant le poids de celles dont la classification ne pose pas de problème.

Les étapes de l'algorithme de boosting :

- Au début, à chaque échantillon de données un poids égal est attribué, l'algorithme utilise ces données non modifiées pour faire les prédictions.
- L'évaluation des prédictions : l'algorithme augmente le poids des échantillons qui présentent une erreur de prédiction importante.
- L'algorithme transmet des données pondérées à l'arbre de décision.
- La réitération des étapes 2 et 3 jusqu'à ce que les erreurs d'apprentissage soient inférieures à un seuil prédéterminé.

Les avantages de la méthode du boosting :

- Efficacité : cet algorithme favorise les caractéristiques augmentant la précision prédictive au moment de son entrainement. C'est utile pour travailler avec de grands jeux des données.
- Facile d'utilisation : il ne nécessite pas de prétraitement des données et facilement interprétable.
- Réduction des biais en améliorant la précision du modèle et ses performances en transformant des apprenants faibles en un apprenant fort.

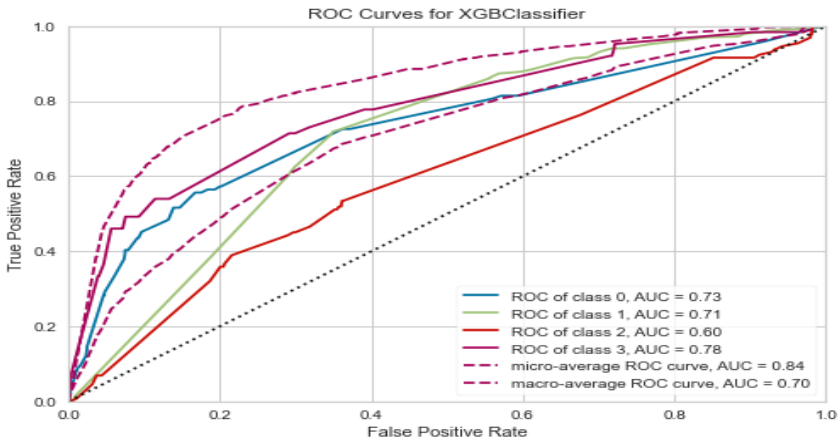


Figure 18 : Courbe ROC AUC (Gradient boosting)

2.3.1.5 **Naive Bayesian**

Le *naive Bayes classifier* est basé sur le **théorème de Bayes**. Ce classifieur est appelé « naïf » cela fait référence à l'existence d'une forte hypothèse : l'indépendance des variables du modèle.

L'idée du théorème de Bayes est de déterminer la probabilité de survenance d'un évènement compte tenu de la probabilité d'un autre évènement qui est déjà survenu. Le théorème de Bayes s'exprime par l'équation suivante :

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{P(B)}$$

Où A et B sont des événements et P(B) ≠ 0.

Désormais, on peut appliquer l'hypothèse d'indépendance.

Maintenant, si deux événements quelconques A et B sont indépendants, alors,

P (A∩ B) = P(A)P(B)

Par conséquent, on peut réécrire l'équation de la manière qui suit :

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Où, y est la variable de classe et X est un vecteur de caractéristiques dépendant (de taille n) où:

$X = (x_1, x_2, x_3, \ldots\ldots, x_n)$, donc

$$P(y|x_1, \ldots, x_n) = \frac{P(x_1|y)P(x_2|y) \ldots P(x_n|y)P(y)}{P(x_1)P(x_2) \ldots P(x_n)}$$

Cette formule peut s'exprimer comme suit :

$$P(y|x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2) \ldots P(x_n)}$$

Puisque le dénominateur reste constant, ce terme peut être supprimé :

$$P(y|x_1, \ldots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

Pour créer un modèle de classification, on détermine la probabilité d'un ensemble donné d'entrées pour toutes les valeurs possibles de la variable de classe y et sélectionne la sortie avec la probabilité maximale.

$$y = argmax_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

L'accuracy est égale à 68.1%, le modèle prédit de manière exacte dans 68,1% des cas classes M et L ont un taux de bonne prédiction faible.

```
Confusion Matrix and Statistics

          Reference
Prediction  XS   S   M   L
        XS  80  30  14  12
        S   46 486 106  37
        M    4   7   8   1
        L    7   6   4  11

Overall Statistics

               Accuracy : 0.681
                 95% CI : (0.6487, 0.7121)
    No Information Rate : 0.6158
    P-Value [Acc > NIR] : 4.102e-05

                  Kappa : 0.3422

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: XS Class: S Class: M Class: L
Precision              0.58824   0.7200 0.400000  0.39286
Recall                 0.58394   0.9187 0.060606  0.18033
F1                     0.58608   0.8073 0.105263  0.24719
Prevalence             0.15949   0.6158 0.153667  0.07101
Detection Rate         0.09313   0.5658 0.009313  0.01281
Detection Prevalence   0.15832   0.7858 0.023283  0.03260
Balanced Accuracy      0.75319   0.6730 0.522050  0.57951
```

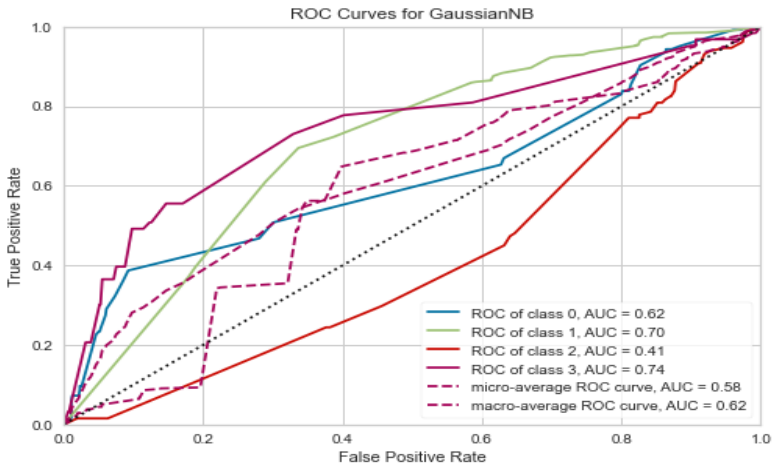


Figure 19 : Résultats et Courbe ROC AUC (naif bayesian)

Le graphe de la courbe ROC confirme le résultat du tableau précédent, la classe 2 a un taux de bonne prédiction trop faible. Il est inférieur à 50% donc non informatif.

Avantages :

- Rapidité pour la classification par rapport aux autres modèles ;
- Il est adapté même pour un petit jeu de données ;

Inconvénients :

- L'hypothèse forte est exigée pour l'utiliser : l'indépendance des variables.

2.3.1.6 **Multinomial logistic regression**

La Multinomial logistic regression est une méthode de classification qui généralise la régression logistique aux problèmes multi-classes. La régression logistique permet d'isoler les effets de chaque variable sur une variable cible. La régression logistique multinomiale est un cas particulier de la régression logistique aux variables qualitatives à trois ou plus modalités.

La régression logistique multinomiale est utilisée pour modéliser les variables nominales, où les probabilités logarithmiques des résultats sont modélisées comme une combinaison linéaire des variables prédictives. Si les catégories d'une variable cible peuvent être ordonnées hiérarchiquement, la distribution est donc multinomiale ordinale.

Soit Y est une variable à expliquer à K modalités et on cherche à modéliser les probabilités suivantes :

$$P(Y_t = j), j = 1 \dots K-1, t = 1, \dots . T$$

L'idée est de choisir une modalité de référence, par exemple la modalité K, et de modéliser les probabilités $p_j(x)$ selon la formule :

$$log\left(\frac{p_j(x_t)}{p_K(x_t)}\right) = \beta_{1j}x_{t1} + \cdots + \beta_{pj}x_{tp} = x_t'\beta_j$$

Où

$$\beta j = (\beta_{1j}, \dots, \beta_{pj})$$

Si K=2, on est dans le cas simple du modèle logistique binaire.

Ce modèle comprend p(K-1) des paramètres à estimer.

La précision de ce modèle est de 67.75% (légèrement inférieure au modèle bayésien), alors que kappa est de 0.323. De même que pour le modèle bayésien, la courbe ROC démontre une très mauvaise prédiction pour la classe 2 (<50%).

```
    True
pr    XS   S   M   L
  XS  77  26  15  14
  S   48 494 110  38
  M    3   8   3   1
  L    9   1   4   8
Confusion Matrix and Statistics

          Reference
Prediction  XS   S   M   L
        XS  77  26  15  14
        S   48 494 110  38
        M    3   8   3   1
        L    9   1   4   8

Overall Statistics

               Accuracy : 0.6775
                 95% CI : (0.6451, 0.7087)
    No Information Rate : 0.6158
    P-Value [Acc > NIR] : 9.863e-05

                  Kappa : 0.323

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: XS Class: S Class: M Class: L
Precision              0.58333   0.7159 0.200000 0.363636
Recall                 0.56204   0.9338 0.022727 0.131148
F1                     0.57249   0.8105 0.040816 0.192771
Prevalence             0.15949   0.6158 0.153667 0.071013
Detection Rate         0.08964   0.5751 0.003492 0.009313
Detection Prevalence   0.15367   0.8033 0.017462 0.025611
Balanced Accuracy      0.74293   0.6699 0.503111 0.556802
```

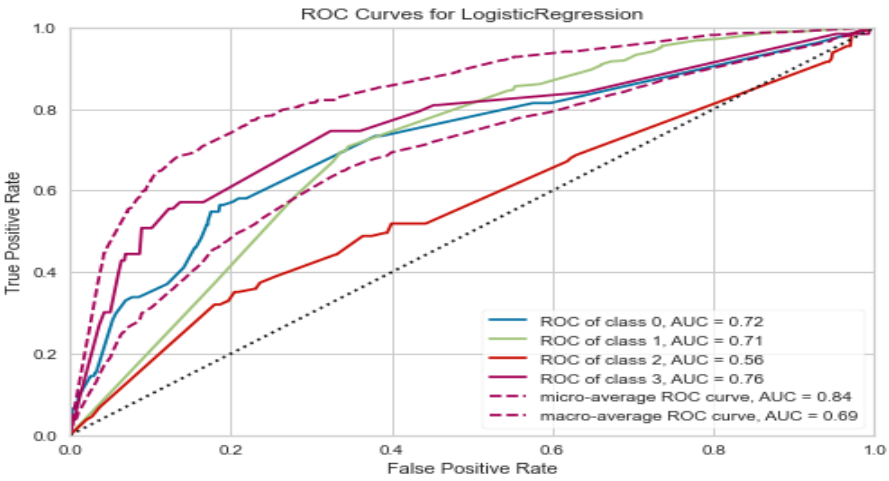


Figure 20 : Résultats et Courbe ROC AUC (GLM)

Dans le cas de l'analyse par les méthodes du Machine Learning, tous les modèles rencontrent le même problème pour la prédiction des sinistres extrêmes. Les modalités de la variable représentant la gravité du sinistre sont déséquilibrées. La classe S est surreprésentée en nombre. L'utilisation de l'accuracy n'est pas la plus adaptée pour ce type de données. La solution possible pour le problème des données déséquilibrées sont des techniques de sous ou sur-échantillonnage. Par exemple, SMOTE (Synthetic Minority Oversampling Technique) est une méthode de suréchantillonnage des observations minoritaires pour leur donner plus d'importance lors de la modélisation.

2.3.1.7 Réseaux de neurones

Le concept de neurone artificiel est l'un des champs de recherche du Deep Learning le plus connu. L'objectif du Deep Learning est le suivant : prédiction d'une sortie Y à partir d'un ensemble de données Xi (input). Un réseau de neurones est constitué de milliers de neurones organisés en couches qui sont interconnectées entre elles par des connexions pondérées (Wi). Chaque nœud est caractérisé par une fonction d'activation Fi et une condition d'activation Ci. L'objectif de la phase d'apprentissage est de trouver une combinaison optimale des Wi/Fi/Ci pour avoir le modèle le plus performant possible.

Pour trouver cette combinaison optimale, on définit une fonction d'erreur qui calcule la différence ente Y et Y'. Où Y est la sortie réelle alors que Y' est la sortie attendue du réseau.

L'erreur de rétropropagation mesure la performance du modèle. Il existe différentes manières de la définir, la plus connue est la MSE (mean squared error) :

$$E(Y, Y') = \frac{1}{n}\sum_{i}^{n}(Y_i - Y'_i)^2$$

A travers cette fonction perte, on cherche à optimiser le modèle (trouver Wi/Ci/Fi) pour minimiser E (Y, Y').

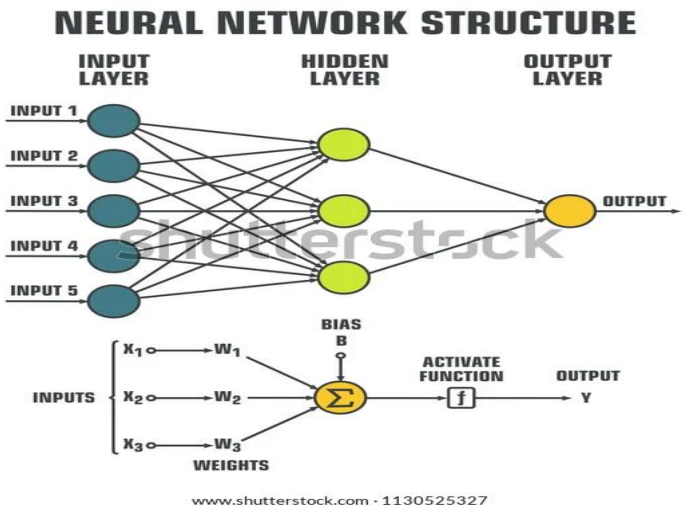


Figure 21 : Neural network structure / Source : www.shutterstock.com

Au début, l'initialisation des poids Wi se fait de manière aléatoire, Y' trouvé permet de calculer E (Y, Y') qu'il faut s'efforcer à diminuer (forward propagation). La valeur de sortie de chaque neurone peut être exprimée par la formule suivante :

$$y_j = b_j + \sum_i x_i w_{ij}$$

La forme matricielle est représentée par la formule suivante :

$$Y = XW + B$$

Où :

$$X = [x_1 \dots . x_i] \quad W = \begin{bmatrix} w_{11} & \cdots & w_{1j} \\ \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{ij} \end{bmatrix} \quad B = [b_1 \dots . b_j]$$

Pour diminuer l'erreur, le paramètre w (poids) doit être modifié.

$$w \leftarrow w - \alpha \frac{\partial E}{\partial w}$$

Où α est le taux d'apprentissage, c'est un paramètre appartenant à l'intervalle [0,1] qu'on peut fixer et $\partial E/\partial W$ est la dérivée de l'erreur par rapport au poids. On doit calculer $\partial E/\partial W$et ajuster les paramètres de chaque couche. Pour le faire, on utilise la règle de dérivation des fonctions composées.

$$\frac{\partial E}{\partial w} = \sum_j \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial w}$$

Donc

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_1} \frac{\partial y_1}{\partial w_{ij}} + \cdots + \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial w_{ij}} = \frac{\partial E}{\partial y_j} x_i$$

$$\frac{\partial E}{\partial B} = \left[\frac{\partial E}{\partial b_1} \frac{\partial E}{\partial b_2} \cdots \frac{\partial E}{\partial b_j} \right]$$

$$\frac{\partial E}{\partial b_j} = \frac{\partial E}{\partial y_1} \frac{\partial y_1}{\partial b_j} + \cdots + \frac{\partial E}{\partial y_j} \frac{\partial y_j}{\partial b_j} = \frac{\partial E}{\partial y_j}$$

Ensuite vient l'étape de la rétropropagation, l'erreur propre à chaque neurone $E(n)$ est calculée. Les poids sont alors mis à jour et la propagation continue jusqu'à atteindre le seuil d'erreur préalablement fixé. Si on a la dérivée de l'erreur par rapport à la sortie $\partial E/\partial Y$, alors on peut calculer la dérivée de l'erreur par rapport aux inputs$\partial E/\partial X$.

$$\frac{\partial E}{\partial X} \leftarrow layer \leftarrow \frac{\partial E}{\partial Y}$$

Où, $\frac{\partial E}{\partial X} = \left[\frac{\partial E}{\partial x_1}, \frac{\partial E}{\partial x_2}, \dots, \frac{\partial E}{\partial x_i} \right]$ et $\frac{\partial E}{\partial Y} = \left[\frac{\partial E}{\partial y_1}, \frac{\partial E}{\partial y_2}, \dots, \frac{\partial E}{\partial y_j} \right]$.

Les couches se succèdent, la sortie d'une couche est l'entrée de la couche suivante. Par conséquent, $\partial E/\partial X$ pour une couche sera $\partial E/\partial Y$ pour la couche précédente. En utilisant

∂E/∂Y, l'ajustement des paramètres de la couche précédente devient possible. Pour calculer ∂E/∂X on utilise de la même manière la règle de dérivation des fonctions composées :

$$\frac{\partial E}{\partial x_i} = \sum_j \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial x_i}$$

Donc

$$\frac{\partial E}{\partial x_i} = \frac{\partial E}{\partial y_1}\frac{\partial y_1}{\partial x_i} + \cdots + \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial x_i} = \frac{\partial E}{\partial y_1}w_{i1} + \cdots + \frac{\partial E}{\partial y_j}w_{ij}$$

Les formules de la couche sont donc :

$$\frac{\partial E}{\partial X} = \frac{\partial E}{\partial Y}W^t \qquad \frac{\partial E}{\partial W} = X^t\frac{\partial E}{\partial Y} \qquad \frac{\partial E}{\partial B} = \frac{\partial E}{\partial Y}$$

Ces calculs étaient linéaires, il est nécessaire d'appliquer également des fonctions non linéaires.

$$\begin{aligned}\frac{\partial E}{\partial X} &= \left[\frac{\partial E}{\partial x_1} \ \cdots \ \frac{\partial E}{\partial x_i}\right] = \left[\frac{\partial E}{\partial y_1}\frac{\partial y_1}{\partial x_1} \ \cdots \ \frac{\partial E}{\partial y_i}\frac{\partial y_i}{\partial x_i}\right] \\ &= \left[\frac{\partial E}{\partial y_1}f'(x_1) \ \ldots \ \frac{\partial E}{\partial y_i}f'(x_i)\right] \\ &= \left[\frac{\partial E}{\partial y_1} \ \ldots \ \frac{\partial E}{\partial y_i}\right] \odot [f'(x_1) \ \ldots \ f'(x_i)] \\ &= \frac{\partial E}{\partial Y} \odot f'(X)\end{aligned}$$

- Apprentissage profond pour le traitement du langage naturel (Natural Language Processing)

Le NLP est une branche de l'intelligence artificielle permettant aux programmes informatiques de comprendre le langage humain. Elle associe l'intelligence artificielle au traitement linguistique. Les applications du NLP sont très nombreuses : traduction automatique, reconnaissance des images, classification de texte, reconnaissance vocale, vérification orthographique, …

La classification de texte consiste à associer une étiquette à un texte, c'est-à-dire à l'associer à une classe bien définie. Les méthodes du deep learning se sont révélées très efficaces pour différents types de données, dont les données de nature textuelle. La base PRC possède une variable textuelle : « Description Of Incident » qui donne la description de l'incident qui s'est produit. La classification des données textuelles vise à identifier le meilleur algorithme de

classification et de définir les meilleures caractéristiques pour ces classifieurs pour bien représenter les données.

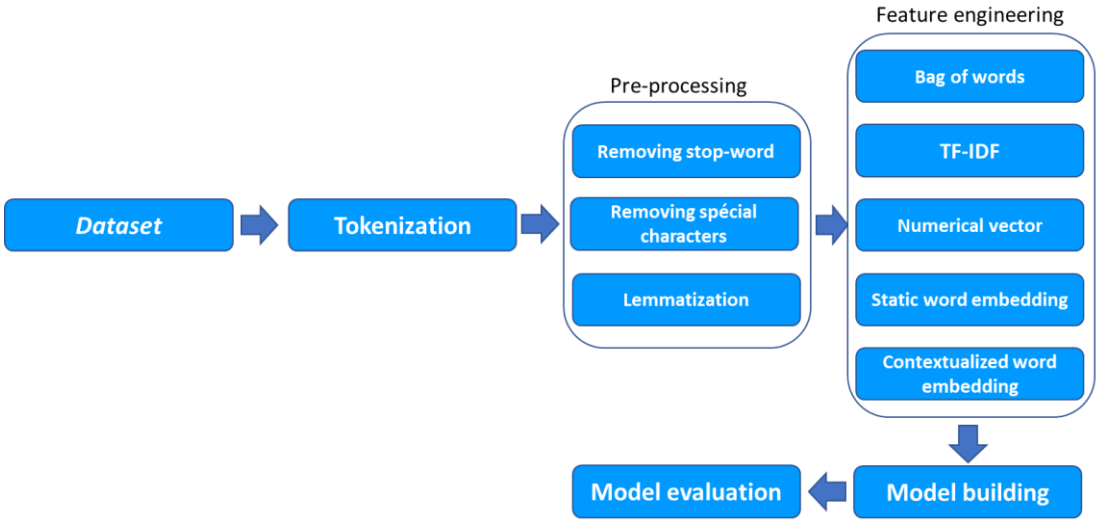


Figure 22 : Classification automatique de textes par les réseaux de neurones

Le processus d'analyse automatique des textes nécessite que ces derniers soient préalablement prétraités, il existe plusieurs méthodes pour le faire.

Le texte doit être nettoyé avant de construire un réseau de neurones : suppression des mots vides (articles, prépositions, pronoms, ...), des caractères spéciaux. La normalisation a pour objectif de transformer tous les mots en minuscule et de supprimer les informations inutiles. La lemmatisation a pour objectif de réduire un mot à sa forme la plus basique possible (la racine) et de les regrouper entre les mots qui ont la même racine. Ce procédé permet de réduire la taille du vocabulaire utilisé et donc augmenter la vitesse du modèle.

- Bag of words (sacs de mots)

Cette approche ne tient pas compte de la sémantique (sens des mots), il crée une matrice d'occurrence des mots, pour chaque mot dans le document elle donne son nombre d'occurrences et élimine toute l'information concernant l'ordre des mots dans le document. Par conséquent les documents similaires sont les documents ayant le contenu similaire.

- TF- IDF permet, à partir de l'ensemble de texte, de connaître l'importance relative de chaque mot dans un document ;

- Word embedding

Le procédé de Word embedding a pour objectif d'associer les mots par rapport à leur sens. Il s'agit d'un réseau de neurone récursif (RNN). Le réseau de neurones de type Word Embeddings va analyser chaque mot en fonction du ou des mots précédents. Il va créer des liens entre des mots qui ont des caractéristiques communes. Un mot sera représenté par un pourcentage d'appartenance aux « classes » créées par le réseau de neurones. Deux mots sont considérés

comme similaires par l'algorithme si leurs représentations vectorielles sont proches dans le même espace continu. Les premières solutions de ce type sont GLOVE, word2vec, etc.

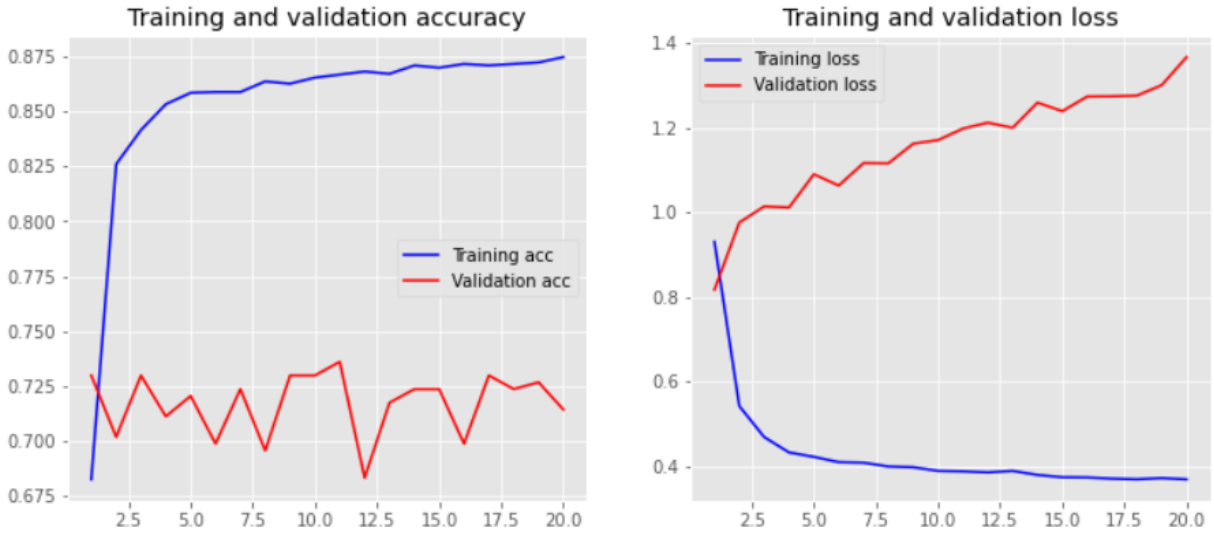


Figure 23 : Training and validation accuracy and loss

Le modèle utilise les réseaux de neurones CNN (Convolutional Neural Network) combiné aux LSTM (Long Short Term Memory networks). Les réseaux CNN modélisent les données en utilisant un mécanisme composé de filtres et de couches de pooling, alors que les LSTM qui permettent de pallier le problème de mémoire courte des réseaux de neurones récurrents (RNN). Le LSTM apprend ce qu'il faut stocker dans le mémoire à long terme et ce qu'il faut oublier.

La précision trouvée est de 0.705.

Ce modèle semble être le plus performant parmi les modèles du Machine Learning testés.

En utilisant les réseaux de neurones dans la prédiction de la variable Type.of.breach, c'est-à-dire le type de violation de données (qui possède 8 modalités :HACK, INSD,CARD,PHYS,PORT,STAT,DISC,UNKN),le résultat obtenu est bien supérieur à celui de la prédiction de la gravité de l'incident cyber. L'accuracy trouvée est égale à 0.8611.

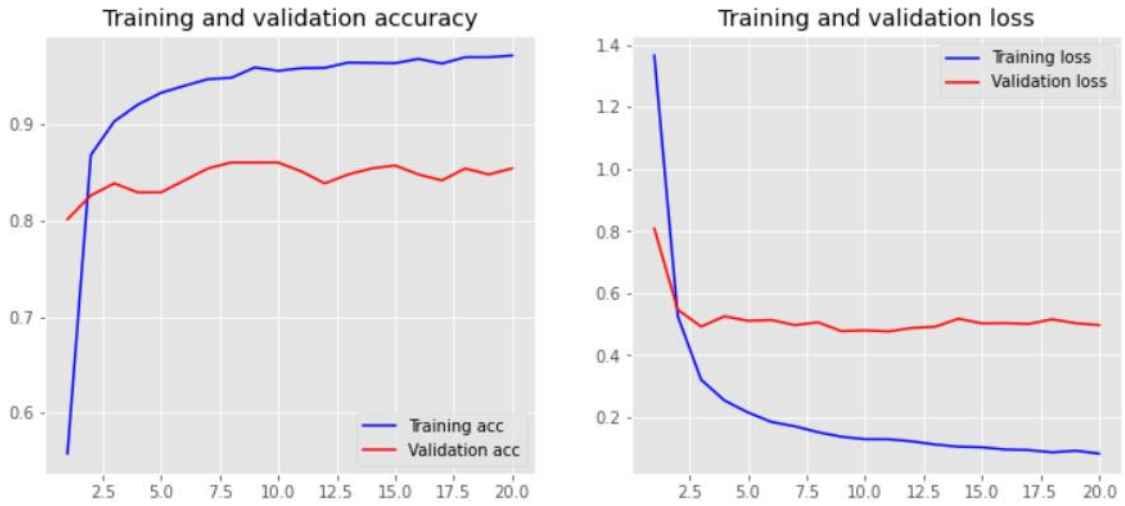


Figure 24 : Training and validation accuracy and loss : (variable Type de la fuite)

Modèles	Accuracy	Kappa
Arbres de décision	0.661	0.327
Random Forest	0.669	0.323
XGBoost	0.671	0.329
Multinomial logistic regression	0.677	0.323
Naive Bayesian	0.681	0.342
Neural Network	0.705	0.351

Figure 25 : Tableau récapitulatif des résultats des modèles du ML

Nous pouvons observer qu'en considérant l'indicateur ROC AUC, les modèles qui semblent les plus satisfaisants sont Random forest et XGboost. Cependant le modèle le plus performant reste le modèle de réseaux de neurones, d'autant que ce modèle-ci n'est pas autant concerné par le problème de données déséquilibrées, car il n'utilise que la variable textuelle (description de l'incident cyber) dans la prédiction de la gravité de l'incident cyber.

2.3.2 Méthodes de régression

Cette section se concentre sur les algorithmes de Machine Learning utilisés à des fins de régression, dans l'objectif de trouver l'algorithme le plus performant pour la prédiction de la gravité d'un sinistre cyber log transformée, ainsi que pour la prédiction du nombre d'incidents survenus.

Chaque modèle a été brièvement présenté dans la section précédente. Les méthodes de Machine Learning sont principalement utilisées pour les problèmes de classification.

Certaines méthodes peuvent être utilisés dans le cas de la régression. Comme pour la classification, les résultats de tous les modèles seront comparés afin de sélectionner le modèle le plus performant au sens des critères de validation choisi : RMSE (écart quadratique moyen). C'est un bon indicateur de la dispersion de la qualité de la prédiction. Il exagère les erreurs importantes, ce qui peut être utile pour éliminer les méthodes entrainant des erreurs importantes.

$$E(Y, Y') = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - Y'_i)^2}$$

En agrégeant nos données en données mensuelles, on n'aura plus d'information sur les caractéristiques de ces sinistres tels que le type de violation, le type d'organisation ou la localisation de l'entreprise.

Nous avons donc préalablement divisé la base PRC suivant certaines modalités des variables qualitatives en question. La mémé méthode sera utilisée dans la partie consacrée à l'étude des séries temporelles.

La base PRC recense des incidents dans un total de 50 états des États-Unis.

Cependant, l'analyse descriptive a mis en évidence le fait que certains États ont connu très peu d'incidents (5 pour l'Idaho, 8 pour le Delaware, …), les paramètres estimés ne seraient pas robustes. Par conséquent, pour mener à bien cette modélisation, la réduction de la quantité d'informations est nécessaire, mais tout en prenant en compte l'hétérogénéité spatiale.

La classification selon les zones de risque homogène face au risque cyber s'avère donc essentielle. Le nombre d'enregistrements compromis moyen va caractériser chacune de ces zones.

2.3.2.1 **Regroupement par zones de risque homogène**

Le regroupement de ces zones géographiques sera réalisé par les méthodes d'analyse de données factorielles.

- *ACP : analyse en composantes principales*

La méthode utilisée est l'analyse en composantes principales (ACP). L'objectif de l'ACP est d'étudier les ressemblances ou les différences entre individus et de révéler des profils d'individus pour pouvoir construire des zones homogènes face au risque cyber.

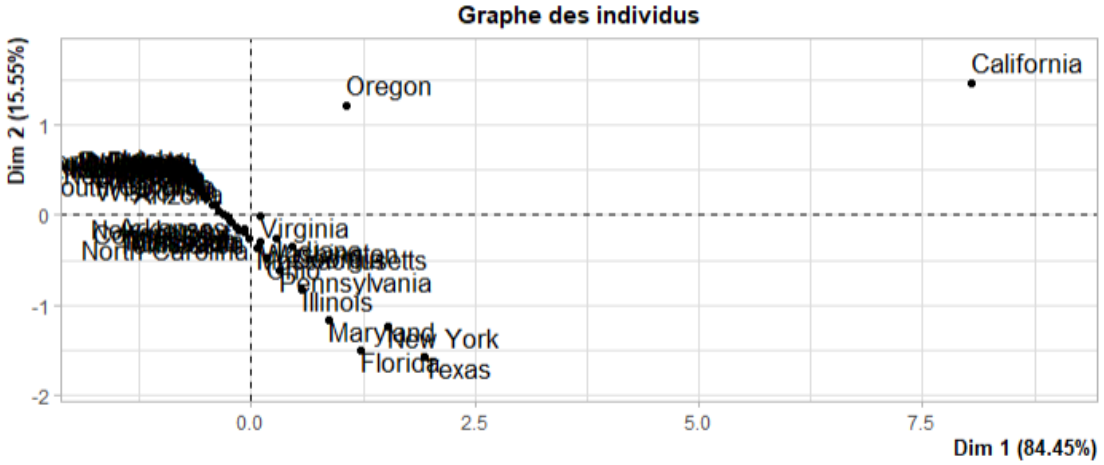


Figure 26 : Graphe des individus (ACP)

Afin de regrouper les Etats en zones de risque homogène, la classification ascendante hiérarchique (CAH) est ensuite réalisée sur les résultats donnés par l'ACP.

- *CAH : classification ascendante hiérarchique*

La classification ascendante hiérarchique est une méthode de classification non supervisée qui a pour l'objectif de construire une hiérarchie sur les individus et se présente sous la forme d'un arbre permettant de visualiser les distances entre individus et groupes d'individus (dendrogramme).

La CAH nécessite de choisir deux paramètres :

- L'indice d'agrégation entre les classes,
- La distance entre les individus.

Dans cette étude, la distance choisie est la distance euclidienne et l'indice d'agrégation est celui de Ward.

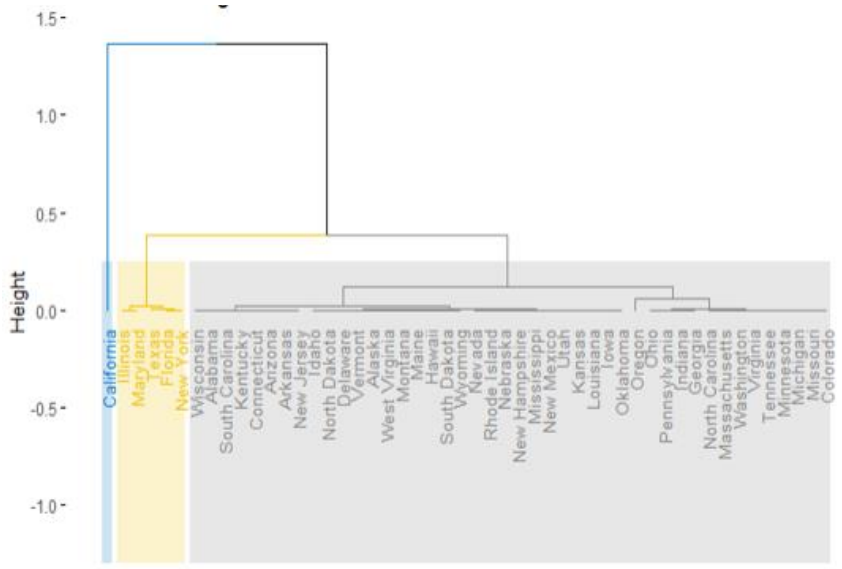


Figure 27 : Cluster dendrogram

Le découpage obtenu est le suivant :

- Cluster 3: California
- Cluster 2: Florida, Texas, Maryland, New York; Illinois.
- Cluster 1 : Tous les états restants

Après avoir terminé cette répartition en zones de risque homogène, on peut estimer ensuite le nombre moyen d'enregistrements compromis pour chaque mois dans chacune de ces zones, en répartissant également les données selon d'autres critères tels que le type de violation et le type d'organisation.

Cependant, diviser cette base de données en un nombre trop conséquent de parties ne serait pas productif, puisque dans ce cas-là, chaque sous-ensemble serait constitué d'un nombre d'observations très faible. Par conséquent, les modèles ne seraient pas suffisamment robustes.

Il serait intéressant de réduire le nombre de modalités des variables qualitatives.

Par exemple, la variable correspondante au type de la violation, qui compte 8 modalités, n'en gardera que 4 :

- HACK,
- ORD (PORT+STAT),
- PHYS,
- INTERN (INSD+DISC).

La modalité CARD n'a pas été conservée, car elle présente trop peu d'observations.

Le même traitement est appliqué à la variable représentant le secteur d'activité de l'entreprise touchée par l'incident. Les 4 modalités restantes sont : ADM (GOV+EDU), retail (BSO+BSR), BSF, MED. La modalité NGO n'a pas été gardée du fait du nombre d'observations trop faible.

Le but de cet apprentissage est de s'entrainer sur un ensemble de données et de pouvoir l'appliquer à de nouvelles données.

Les échantillons utilisés pour l'apprentissage sont bien différents de ceux utilisés pour tester.

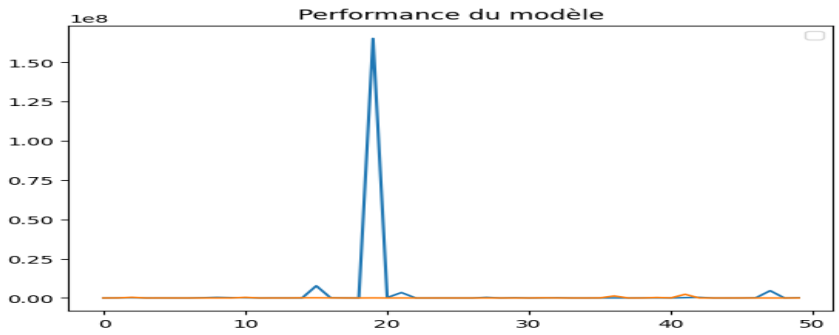
Pour améliorer la performance des modèles, la méthode grid search va être utilisée, comme dans le cas des modèles de classification. Cette méthode permet de tester plusieurs combinaisons de paramètres possibles, en fixant pour chaque hyperparamètre un ensemble de valeurs qu'il peut prendre.

Nous avons entrainé notre modèle en testant chaque combinaison et avons gardé la meilleure.

Une transformation logarithmique a été effectuée sur la variable cible (nombre d'enregistrements compromis), car une grande variabilité existe : les valeurs sont réparties entre 0 et 3 milliards et les études antérieures sur la variable non transformée donnaient des résultats peu probants.

La métrique RMSE était particulièrement élevée (82 millions).

Comme nous pouvons observer sur le graphique, la prédiction est extrêmement basse :



Nous avons aussi cherché à prédire le log transformé de la moyenne de la sévérité mensuelle des sinistres cyber et le nombre d'incidents par mois (fréquence).

Les modèles de régression testés initialement ne donnant pas de résultats probants (R2<0.05), nous avons donc procédé à un retraitement de la base pour pouvoir utiliser dans les modèles de régression les données mensuelles en fonction du secteur et de la zone géographique. Ce procédé a permis d'améliorer sensiblement les performances des modèles. Même les modèles les moins performants donnent un R2 supérieur à 0.58. Dans la section suivante, nous allons présenter les résultats obtenus.

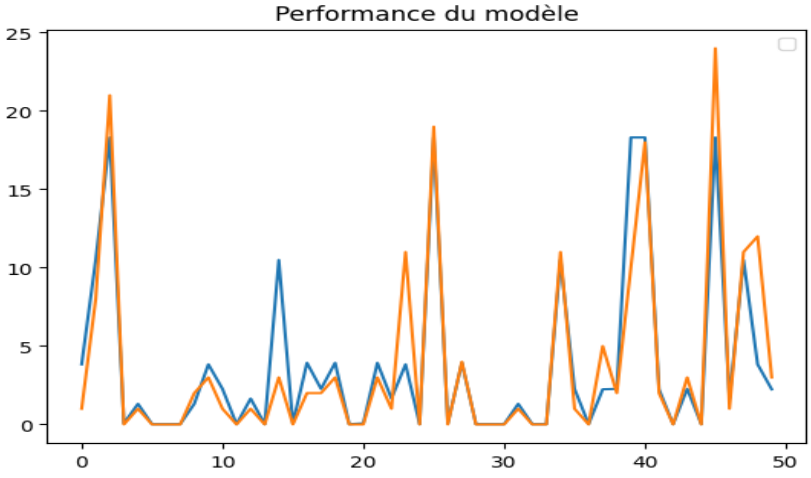
Dans cette section, les algorithmes testés dans cette partie sont les suivants :

- Les Arbres de décision ;
- Le Random Forest ;
- Le Gradient boosting (XGBOOST) ;
- Les Réseaux de neurones.

2.3.2.2 **Arbres de décision**

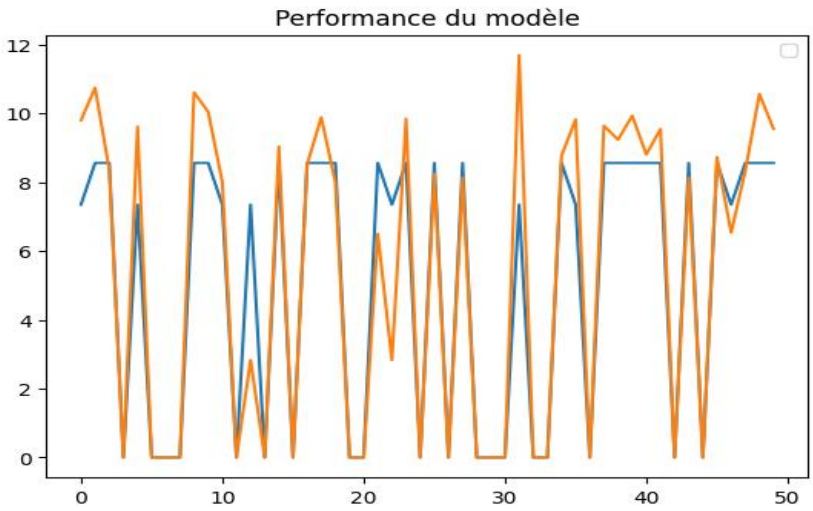
Les modèles de régression utilisés donnent les résultats suivants :

- Prédiction de la fréquence (nombre d'incidents cyber)



NB d'incidents		
	Naif	Grid search
RMSE	4.184	3.152
R2	0.580	0.762

- Prédiction de la gravité (nombre d'enregistrements compromis log transformé)



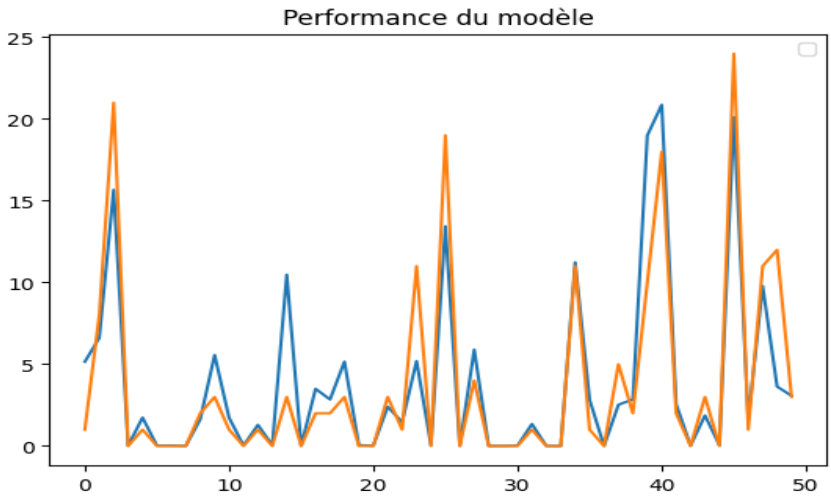
Log(gravité)		
	Naif	Grid search
RMSE	1.590	1.490
R2	0.866	0.882

Figure 28 : Performance des modèles de régression (Arbre de décision)

2.3.2.3 **Random Forest**

Les modèles de régression utilisés donnent les résultats suivants :

- Prédiction de la fréquence (nombre d'incidents cyber)



NB d'incidents		
	Naif	Grid search
RMSE	3.727	3.314
R2	0.667	0.736

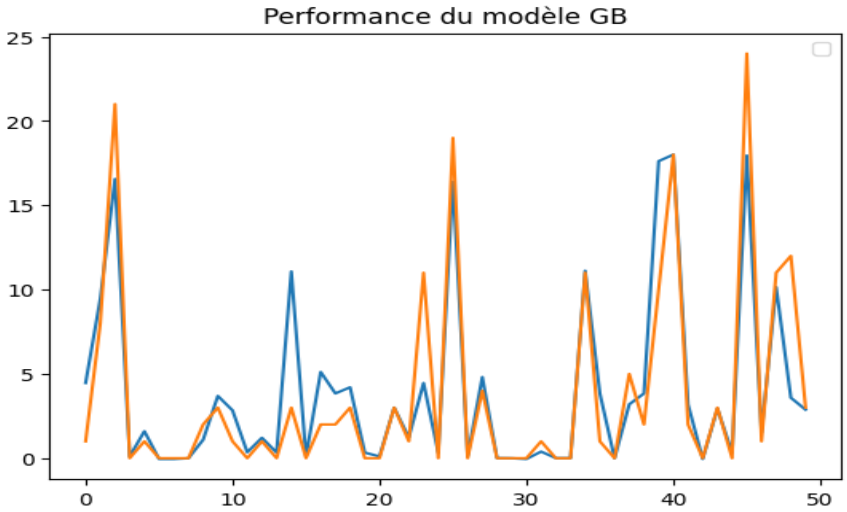
- Prédiction de la gravité (nombre d'enregistrements compromis log transformé)

Log(gravité)		
	Naif	Grid search
RMSE	1.577	1.528
R2	0.868	0.876

Figure 29 : Performance des modèles de régression (Random Forest)

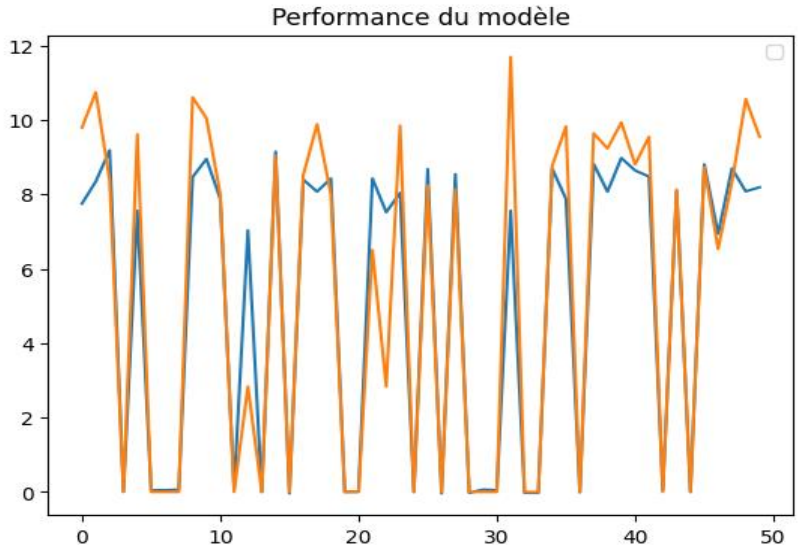
2.3.2.4 **Gradient Boosting**

- Prédiction de la fréquence (nombre d'incidents cyber)



NB d'incidents		
	Naif	Grid search
RMSE	3.359	3.196
R2	0.729	0.755

- Prédiction de la gravité (nombre d'enregistrements compromis log transformé)



Log(gravité)		
	Naif	Grid search
RMSE	1.567	1.484
R2	0.869	0.882

Figure 30 : Performance des modèles de régression (Gradient boosting)

2.3.2.5 **Réseaux de neurones**

Dans cette partie, plusieurs structures de réseaux de neurones ont été testés.

Premièrement, un modèle standard a été testé (« Naïf » dans le tableau ci-dessous).

Deuxièmement, la détermination des paramètres les plus optimaux a été introduite.

Les hyperparamètres à déterminer sont :

- l'architecture du réseau (nombre des couches cachées, nombre de neurones par couche, …),
- le nombre maximum d'itérations,
- la taille des ensembles pris en compte à chaque itération,
- l'erreur maximum tolérée,
- un mode d'estimation de l'erreur,
- une méthode de gradient …

En augmentant le nombre de neurones, on peut provoquer le sur-apprentissage du modèle.

Nous avons utilisé un réseau de neurones constituée de deux couches cachées.

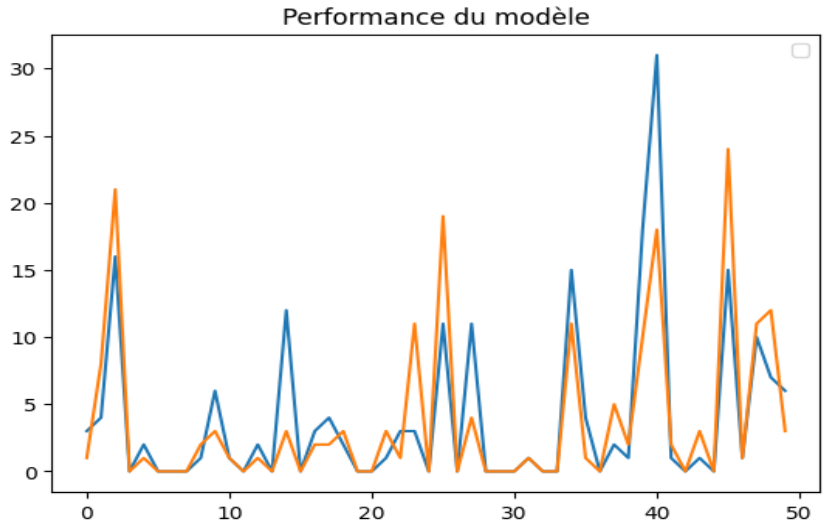
Les deux couches sont composées de 100 neurones, les fonctions d'activation utilisées sont :

- ReLu (rectified linear unit) dont la formule est :$f(x) = max(0, x)$
- tanh, qui est tout simplement une tangente hyperbolique :

$$tanh(x) = \frac{exp(x) - exp(-x)}{exp(x) + exp(-x)}$$

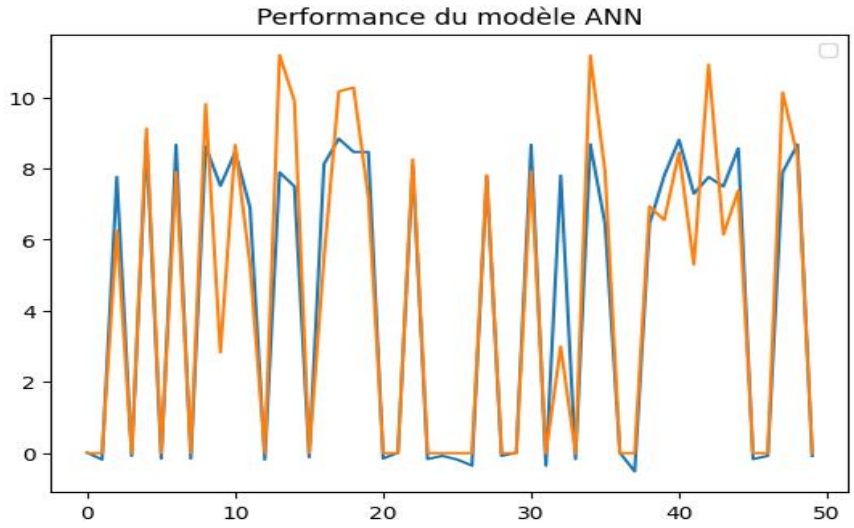
Le but est de comparer les résultats des réseaux de neurones avec les autres modèles de Machine Learning.

- Prédiction de la fréquence (nombre d'incidents cyber)



NB d'incidents		
	Naif	Grid search
RMSE	2.540	2.496

- Prédiction de la gravité (nombre d'enregistrements compromis log transformé)



Log(gravité)		
	Naif	Grid search
RMSE	1.569	1.545

Figure 31 : Performance des modèles de régression (Neural Network)

2.3.2.6 **Tableau récapitulatif**

Pour avoir une meilleure vue d'ensemble des résultats, un tableau résumant les performances des modèles de régression de régression est présenté :

- Selon la métrique RMSE :

Modèle	NB d'incidents		Log(gravité)	
	Naïf	Grid search	Naïf	Grid search
Arbres de décision	4.184	3.152	1.590	1.490
Random Forest	3.727	3.314	1.577	1.528
Gradient Boosting	3.359	3.196	1.567	1.484
Réseaux de neurones	2.540	2.496	1.569	1.545

- Selon la métrique R2 :

Modèle	NB d'incidents		Log(gravité)	
	Naïf	Grid search	Naïf	Grid search
Arbres de décision	0.580	0.762	0.866	0.882
Random Forest	0.667	0.736	0.868	0.876
Gradient Boosting	0.729	0.755	0.869	0.882
Réseaux de neurones	0.750	0.774	0.868	0.873

Figure 32 : Tableaux récapitulatifs (performances des modèles)

Le retraitement de la base a permis d'augmenter sensiblement les performances des modèles.

Le modèle de réseaux de neurones s'est révélé particulièrement performant dans la prédiction du nombre d'incidents mensuels (à savoir pour l'estimation de la fréquence : nombre de sinistres). Concernant la sévérité des incidents, tous les modèles ont montré des performances relativement proches. Cette fois-ci, c'est le modèle de gradient boosting qui est le plus performant.

Nous considérons que la performance des modèles de neurones peut être significativement améliorée en déterminant une meilleure calibration. Cependant en gagnant en précision, on « gagne » également en complexité, car il s'agît d'une boite noire.

Les autres modèles testés présentent l'avantage d'être plus simples et plus rapides à calibrer pour des performances sensiblement équivalentes.

Les modèles de régression donnent des résultats plutôt satisfaisants, comme nous pouvons le constater sur les graphiques représentant la prédiction des variables sur l'échantillon test, mais également la qualité de la prédiction et comparaison des modèles dans les tableaux récapitulatifs ci-dessus.

Nous insistons sur le fait que la variable cible (nombre d'enregistrements corrompus) a subi une transformation logarithmique dans l'objectif d'améliorer les performances des modèles.

Il s'agit d'un point important car la transformation inverse (exponentielle) se traduirait par des écarts importants.

2.4 Mise en place des méthodes des séries temporelles dans la prédiction de la sévérité des incidents cyber

Les données utilisées sur la période de 2010 à 2016, permettront de prédire le nombre moyen d'enregistrements compromis en 2017.

Les notations utilisées sont les suivantes :

- n_i est le nombre des sinistres cyber pour le mois i
- N_j est le nombre des sinistres pour l'année j
- g_i est le nombre d'enregistrements compromis pour le mois i
- G_j est le nombre d'enregistrements compromis pour l'année j
- y_i est le nombre moyen d'enregistrements compromis pour le mois i

$$y_i = \frac{g_i}{n_i}$$

- Y_j est le nombre moyen d'enregistrements compromis pour l'année j

$$Y_j = \frac{G_j}{N_j}$$

Où $i \in [|1 ; 96 |]$ et j $\in$ [| 2010 ; 2017 |]

En agrégeant nos données en données mensuelles, nous n'aurons plus d'information sur les caractéristiques de ces sinistres tels que le type de violation, le type d'organisation ou la localisation de l'entreprise.

Nous avons donc préalablement divisé la base PRC suivant certaines modalités des variables qualitatives en question.

La classification selon les zones de risque homogène face au risque cyber s'avère donc essentielle comme pour les modèles de régression vus dans la section précédente.

2.4.1 Modèles théoriques des séries temporelles

Dans cette partie, nous nous intéressons à différents modèles théoriques.

Une série temporelle est une série de données qui est indexée par le temps. En analyse de séries temporelles, c'est le temps qui est une variable explicative.

L'analyse et la prédiction de ces séries présente un certain intérêt spécialement pour certaines industries ou secteurs d'activités.

Une série temporelle est composée de ces trois éléments :

- Tendance (ou trend T_t) est une évolution d'une série à long terme.
- Saisonnalité (S_t) est une présence d'une structure périodique. C'est donc un processus se répétant à intervalles de temps réguliers. Cette composante saisonnière est *p*-périodique, c'est-à-dire $S_{k+p} = S_k$, pour tout *k*.

- Erreur (e_t) (ou composante aléatoire) est une partie de la série temporelle inexplicable par une tendance et une saisonnalité. Si le résidu n'est pas stationnaire, c'est-à-dire les résidus évoluent dans le temps (moyenne et variance ne sont pas constantes), alors certaines composantes du modèle ne sont pas bien expliquées par le modèle.

Nous pouvons donc décomposer une série temporelle par

- $X_t = T_t + S_t + e_t$ Si la série temporelle est additive ;
- $X_t = T_t * S_t * e_t$ Si la série temporelle est multiplicative.

2.4.1.1 Modèle ETS (Exponential smoothing)

Nous utilisons le lissage exponentiel afin de faire une prévision de données chronologiques.

Le lissage exponentiel est un moyen qui permet d'analyser les données temporelles en accordant plus d'importance aux données récentes par rapport aux données anciennes.

Cette méthode permet de mieux faire ressortir et visualiser les tendances en matière de projection de la sévérité des futurs incidents.

L'abréviation ETS signifie Error Trend Seasonal.

Ce sont ces trois composantes principales. Il y a trois modalités pour chaque terme : N (aucun), A (additif) et M (multiplicatif). Par exemple, ETS (M, N, N) est un lissage exponentiel simple avec des erreurs multiplicatives.

Il permet donc tester des tendances et des saisonnalités de différents types, ce qui le rend plutôt flexible.

Pour pouvoir sélectionner la meilleure combinaison entre ces 3 modalités, les critères tels que AIC (Akaike's Information Criterion) ou BIC (Bayesian Information Criterion) sont très utiles.

Il y a trois types de lissage exponentiel :

- Lissage exponentiel simple peut être utilisé pour des séries temporelles qui ne présentent ni tendance ni saisonnalité
- Lissage exponentiel double s'applique aux séries chronologiques présentant une composante de tendance
- Lissage exponentiel triple prend en charge et la tendance et la saisonnalité.

La formule utilisée pour le calcul d'une méthode de lissage exponentiel est la suivante :

$$S_{t+1} = \beta X_t + (1 - \beta) S_t$$

Dans cette formule, les observations les plus récentes sont pondérées par le coefficient β et les prévisions plus anciennes par le coefficient $1 - \beta$.

Nous pouvons développer les calculs :

$$S_{t+1} = \beta X_t + (1 - \beta)(\beta X_{t-1} + (1 - \beta) S_{t-1})$$

… …

$$S_{t+1} = \beta x_t + \beta(1-\beta)X_{t-1} + \beta(1-\beta)^2 X_{t-2} + \beta(1-\beta)^3 X_{t-3} + \cdots$$

L'étape essentielle pour le lissage exponentiel est de bien choisir la valeur de la constante de lissage β.

Avantages du lissage exponentiel :

- Simplicité des calculs ;
- Il fonctionne même avec un petit nombre d'observations ;
- La qualité des prédictions.

2.4.1.2 Modèle ARIMA (Autoregressive Integrated Moving Average Model)

Le nom ARIMA (p, d, q) est composée de trois fonctions AR, MA et I.

L'AR signifie « autorégressive », c'est-à-dire il prédit des valeurs futures en se fondant sur les valeurs passées du processus. Dans l'AR (p), p représente le nombre d'observations décalées.

La MA (Moving Average) est la moyenne mobile, elle représente la valeur moyenne du processus sur une période donnée.

Elle est mobile, car elle est recalculée à chaque fois qu'une nouvelle observation est ajoutée. De cette manière les valeurs extrêmes sont lissées permettant d'avoir une tendance moyenne sur le long terme. Dans le modèle $MA(q)$, q représente la taille de la fenêtre de moyenne mobile.

En différenciant les séries temporelles, on peut les rendre stationnaires. Le nombre de différentiations est représenté par le paramètre d.

Le modèle $ARIMA(p, d, q)$ est donné par la formule suivante :

$\phi(L)(1-L)^d X_t = \mu + \theta(L)\varepsilon_t$, $\forall t \geq 0$ et ε_t est un bruit blanc de variance σ^2.

Où $\quad \phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$ avec $\phi_p \neq 0$

$\theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$ avec $\theta_q \neq 0$

Une transformation logarithmique a été effectuée sur la variable cible (nombre d'enregistrements compromis) car une grande variabilité existe : les valeurs sont réparties entre 0 et 3 milliards.

Pour commencer, l'analyse d'une série entière sera faite, avant de considérer chaque zone géographique séparément.

- **Base sans distinction de zone**

La série temporelle étudiée sans distinguer les zones géographiques :

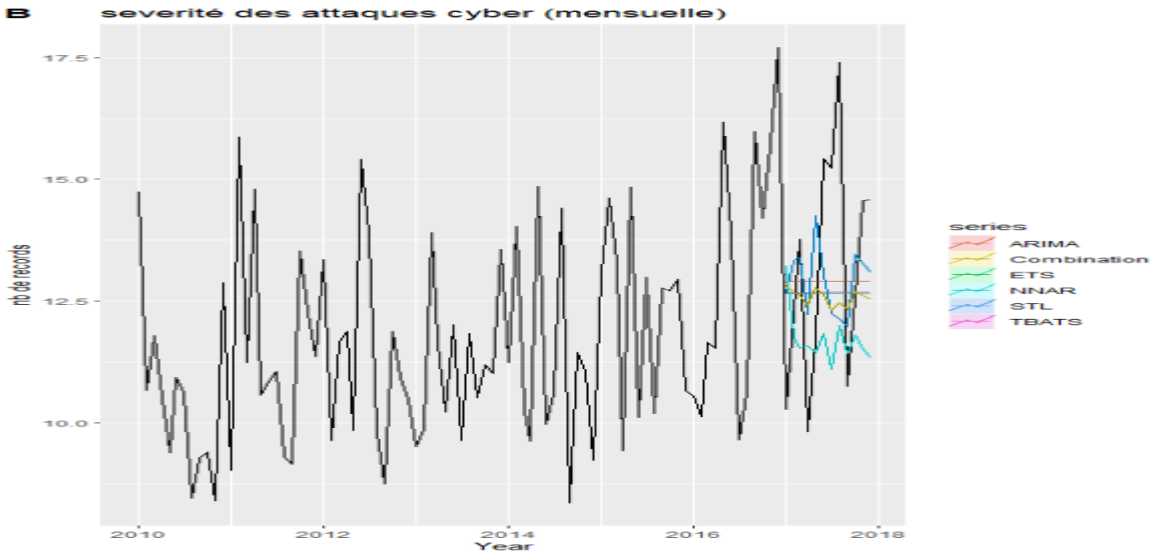
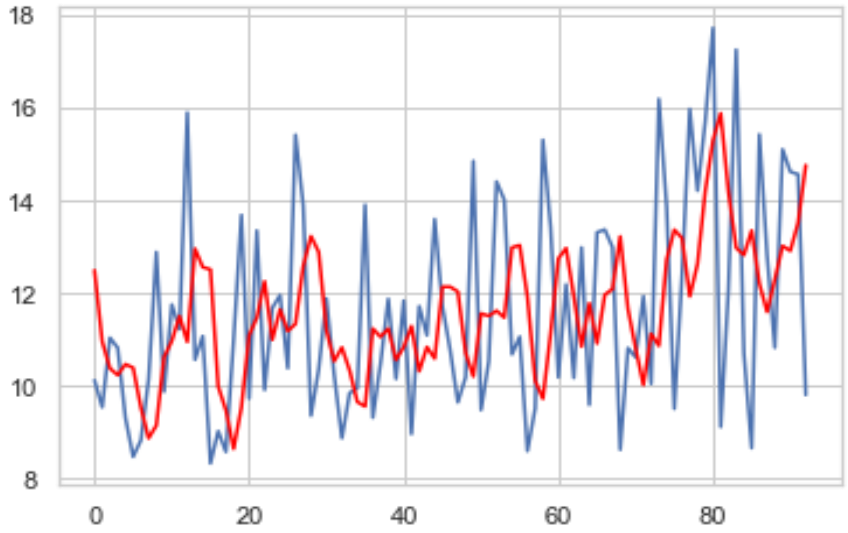


Figure 33: Zone sans distinction des zones géo (performances)



Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA (3)
RMSE	2.52	2.47	2.55	3.15	2.66	2.76	2,49

En comparant les résultats obtenus par zone géographique avec une série initiale (sans distinction des zones), on trouve que la prédiction est améliorée pour la zone 1 incluant presque 44 états, tandis qu'elle baisse pour le cluster 3 (California).

- Fréquence de la base sans distinction de zone géographique

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	18.62	12.89	21.53	15.77	19.73	17.50

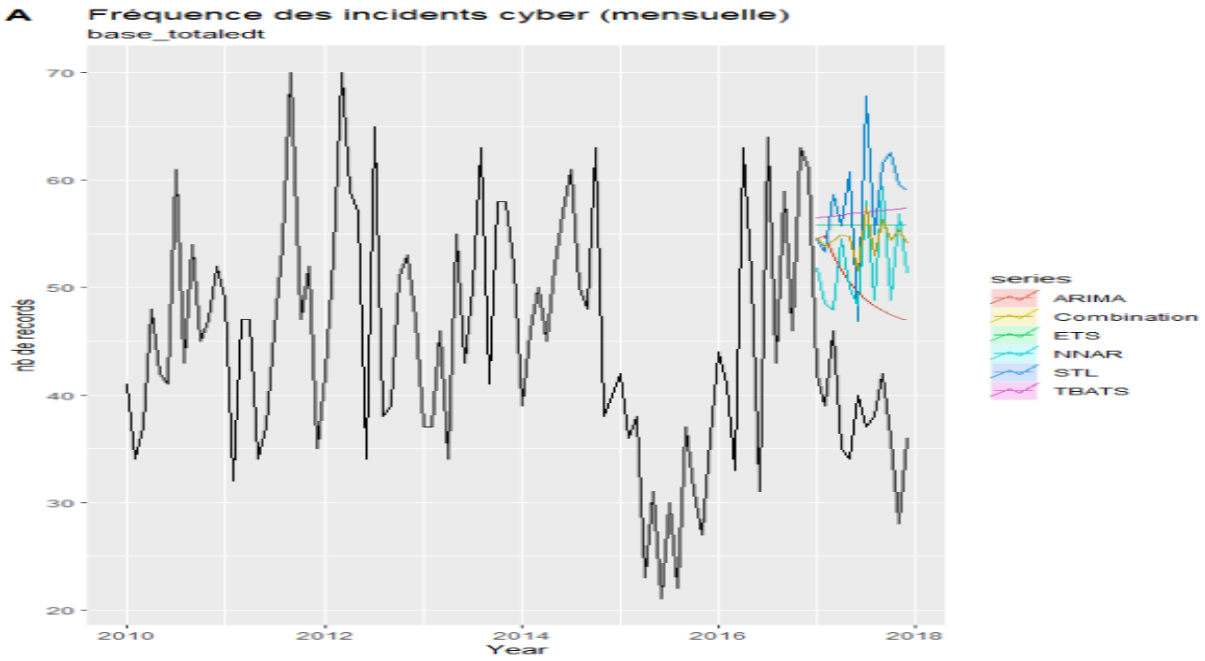
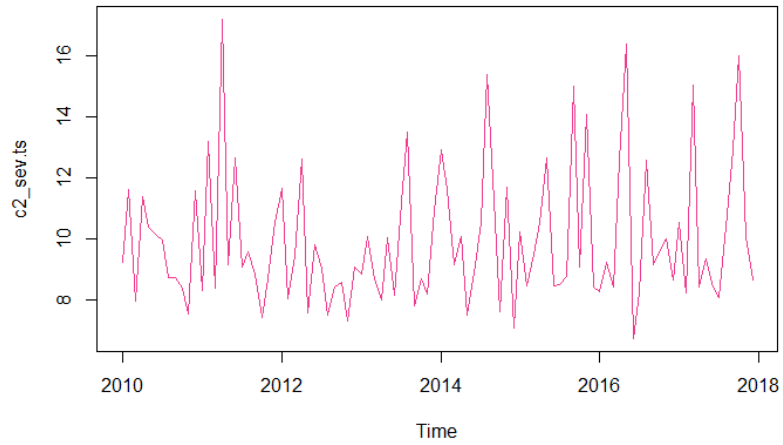


Figure 34 : ST de la fréquence (zone entière)

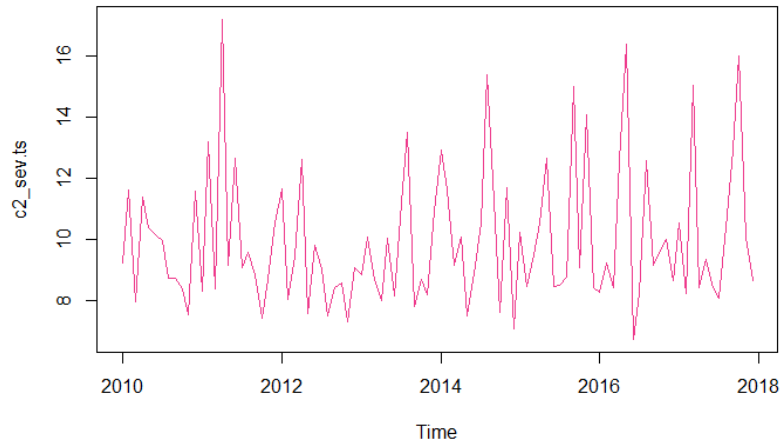
Le modèle le plus performant pour la prédiction de la fréquence (nombre d'incidents mensuels) est aussi le modèle ARIMA.

Ensuite nous allons étudier les séries en distinguant les zones géographiques pour pouvoir améliorer la performance. Nous avons trois zones géographiques, pour chaque zone l'analyse sera faite.

Zone géographique 1



Zone géographique 2



Zone géographique 3

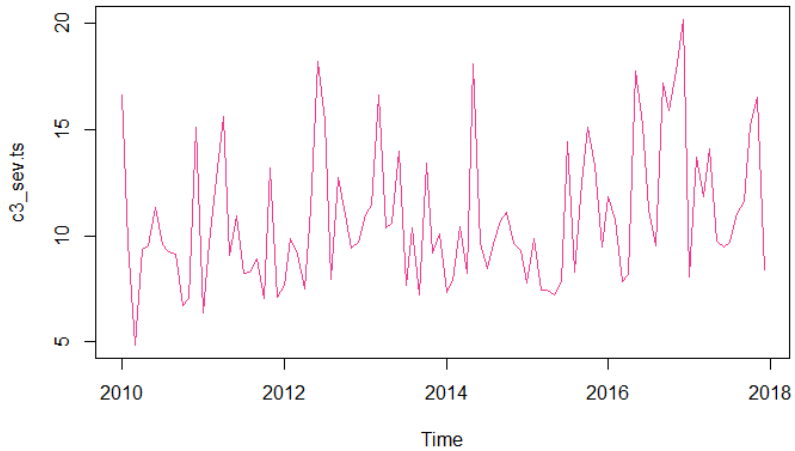
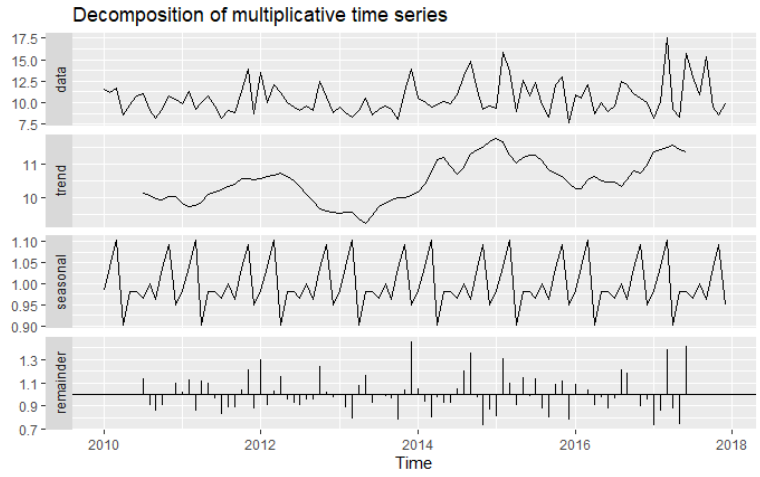


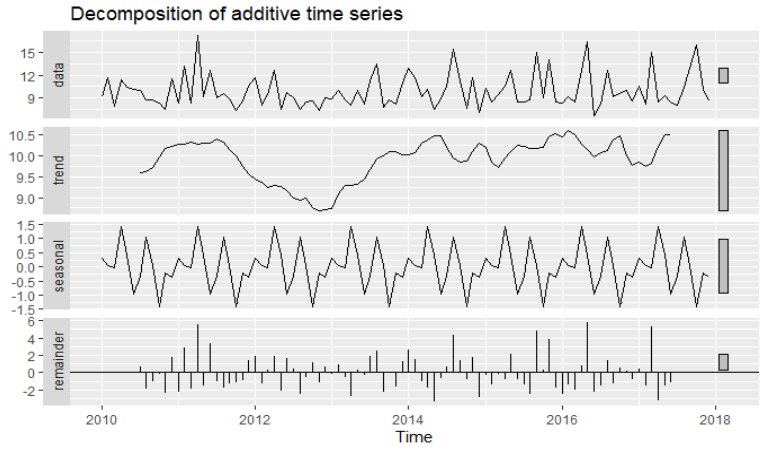
Figure 35 : ST des 3 zones géographiques

Les décompositions de ces séries sont présentées sur le graphe suivant :

Zone géographique 1



Zone géographique 2



Zone géographique 3

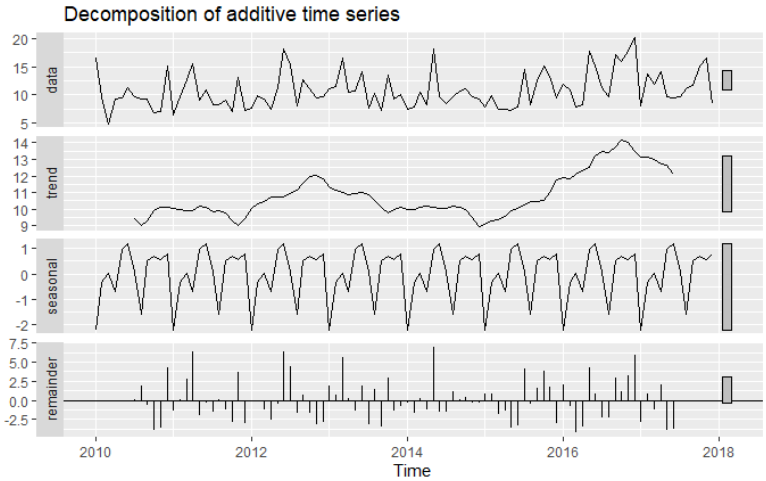


Figure 36 : Décomposition des séries temporelles des 3 zones géographiques

Pour mieux examiner les caractéristiques de ces séries temporelles, les fonctions d'autocorrélation (Autocorrelation Function – ACF) et de l'autocorrélation partielle (Partial autocorrelation Function – PACF) ont été réalisées, sur les trois zones géographiques.

L'autocorrélation est un phénomène de corrélation des valeurs actuelles aux valeurs précédentes ou suivantes.

Les pics au-delà de 0 ne dépassent pas les lignes bleues.

Les corrélogrammes ci-dessous montrent une autocorrélation nulle pour les 3 zones. Il y a une seule autocorrélation non nulle est celle de décalage de 0.

Les séries semblent donc aléatoires.

Il est difficile de modéliser le bruit blanc, car nous ne sommes pas en mesure d'extraire les paramètres des modèles à partir des corrélogrammes ACF et PACF.

Zone géographique 1

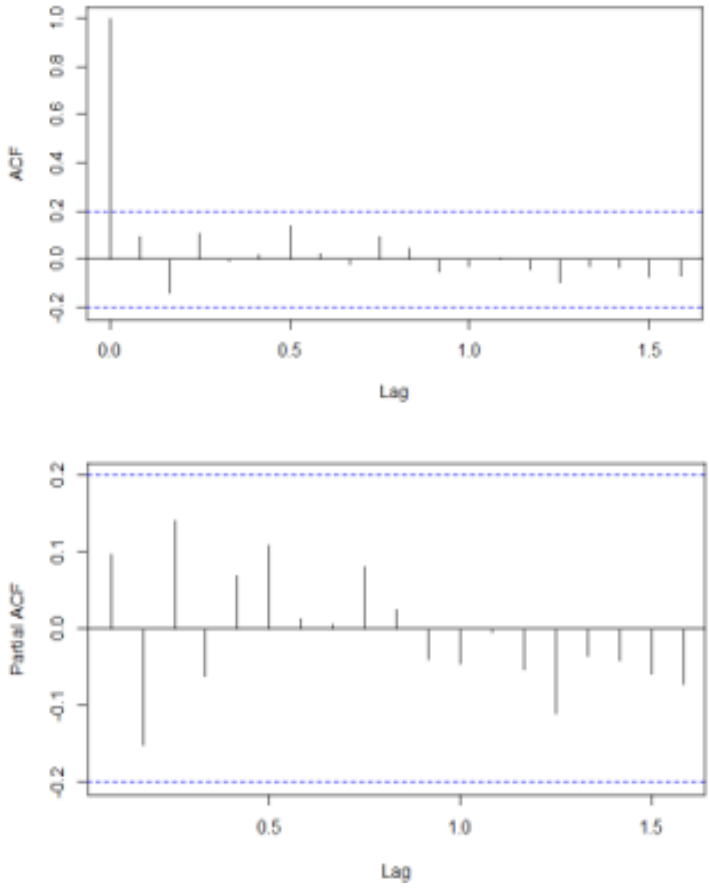


Figure 37 : Les graphes des ACF et PACF – Zone 1

Zone géographique 2

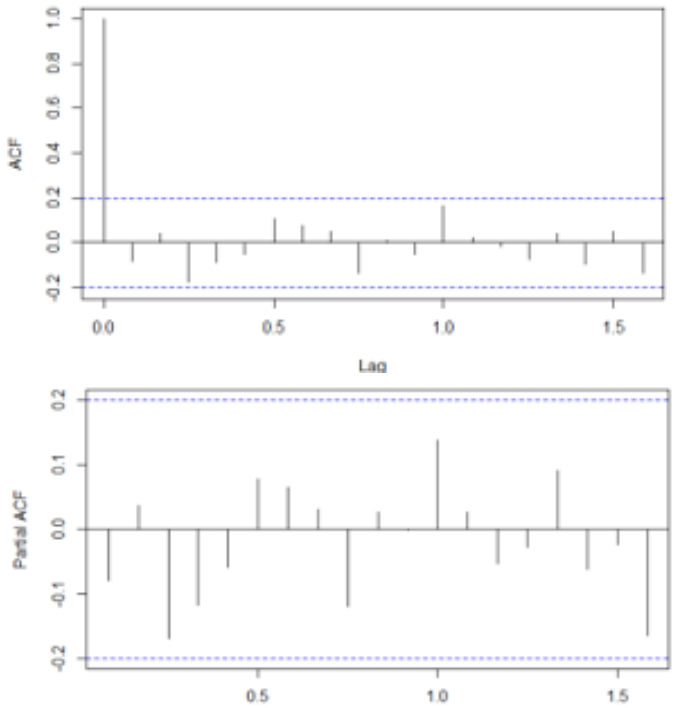


Figure 38 : Les graphes des ACF et PACF – Zone 2

Zone géographique 3

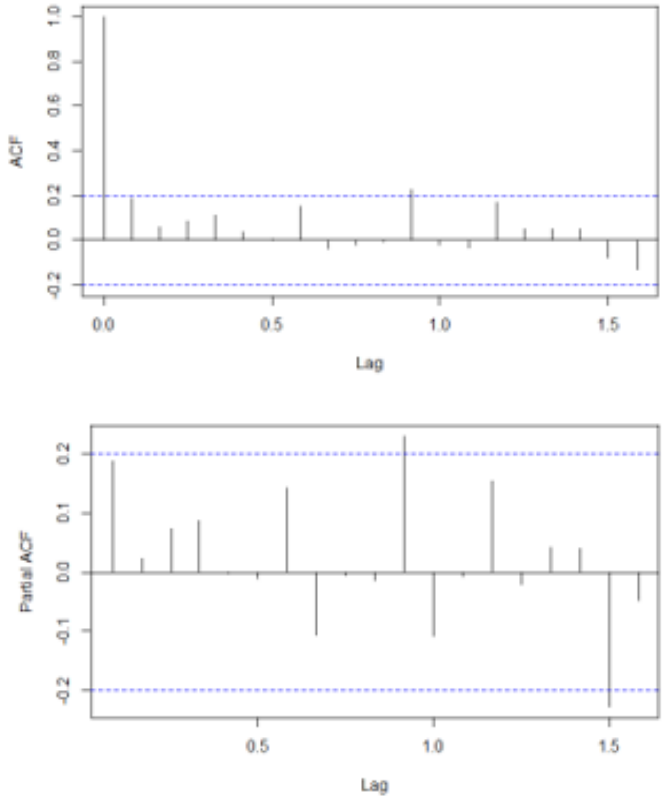


Figure 39 : Les graphes des ACF et PACF – Zone 3

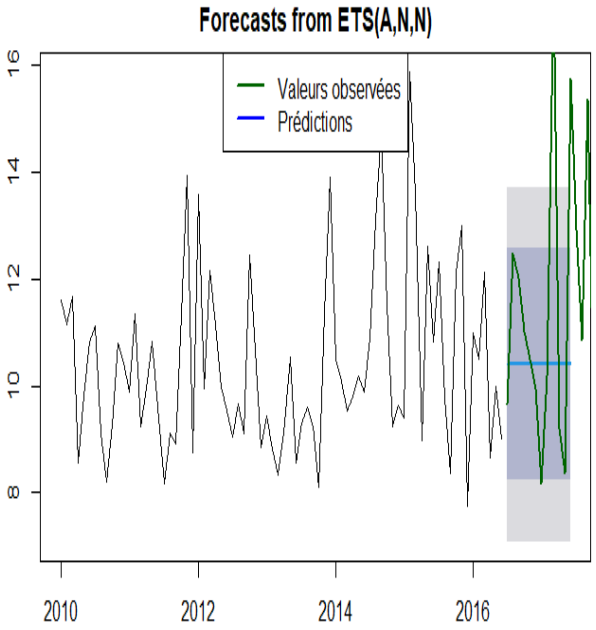
Prédiction par les modèles ETS et MA:

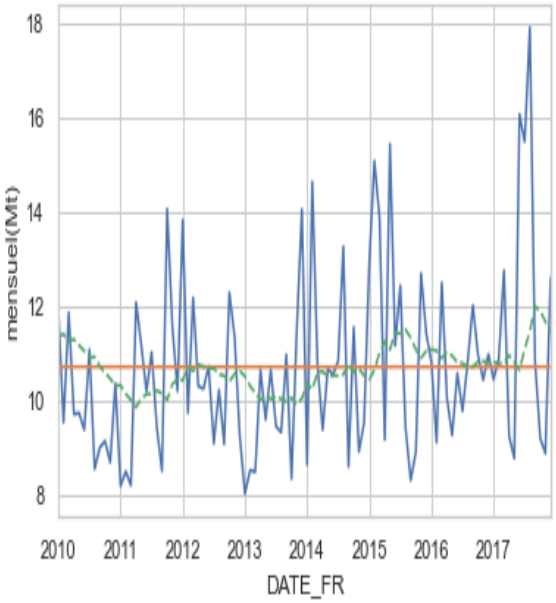
Le modèle ETS choisi est le celui qui minimise l'AIC est le modèle ANN.

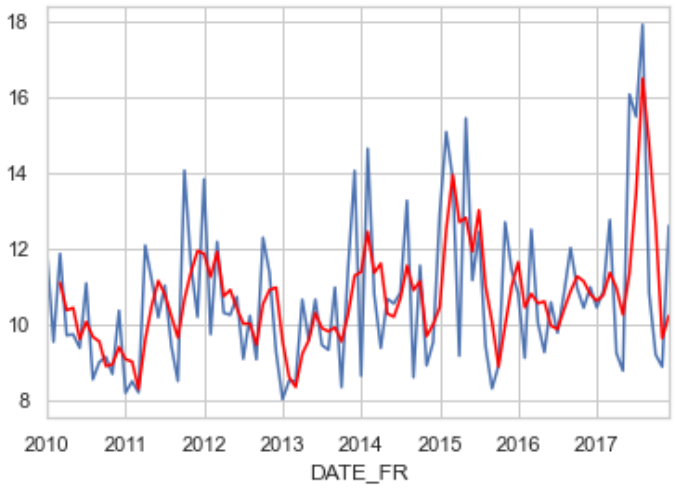
Les prédictions obtenues sont représentées sur les graphes suivants.

Zone géographique 1

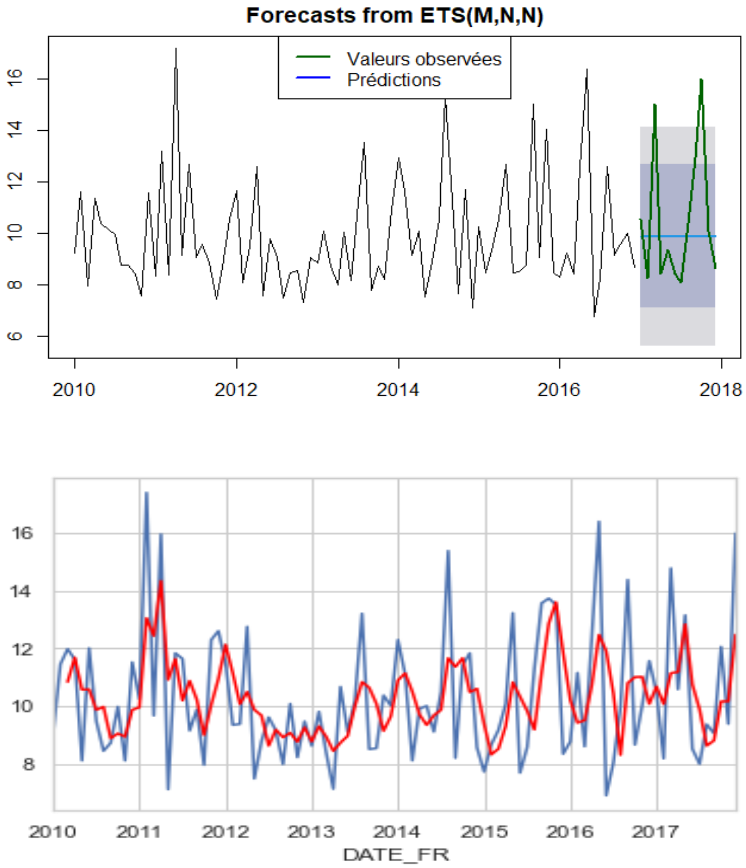
Forecast MA – zone 1





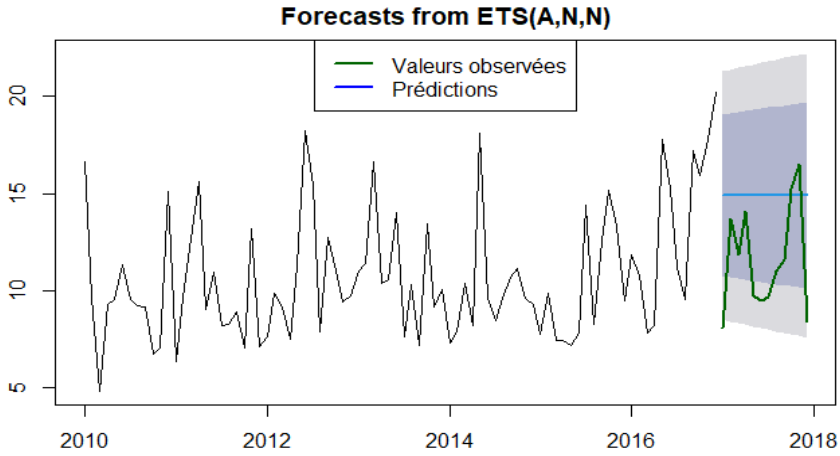


Zone géographique 2



Forecast MA(3) -zone 2

Zone géographique 3



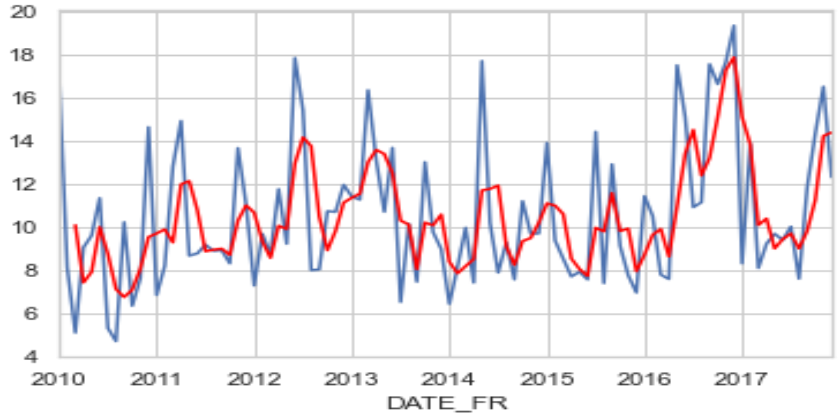
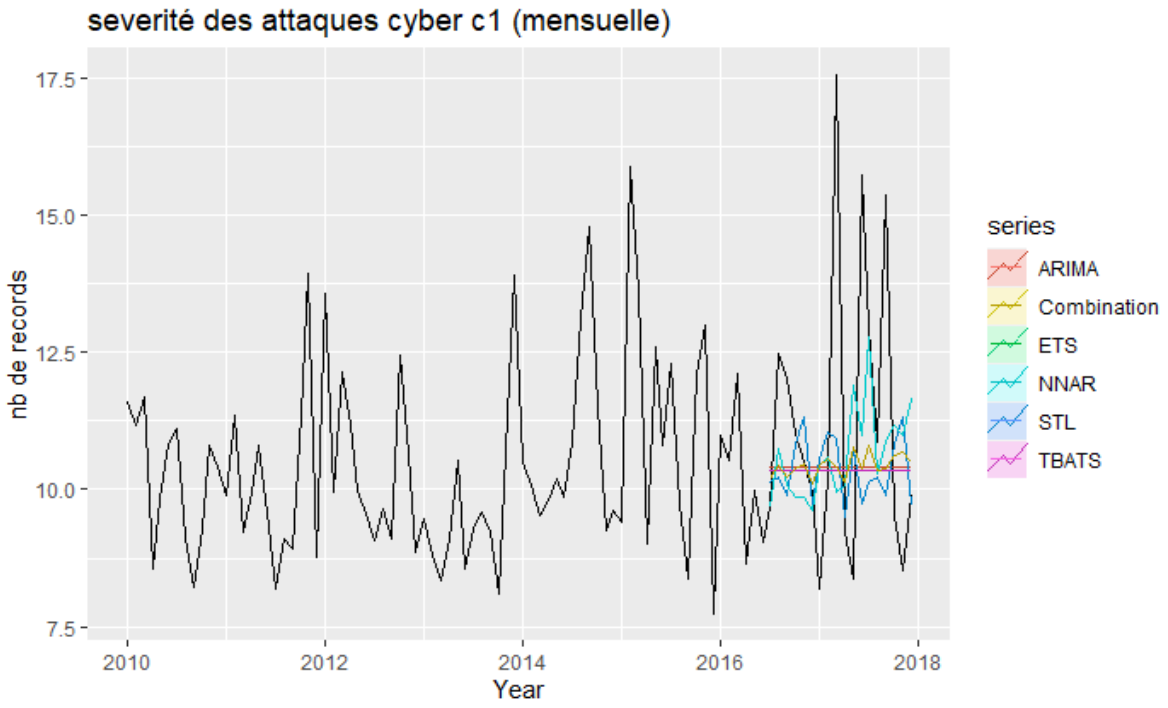


Figure 40 : Prédictions de la sévérité par les modèles ETS et MA - Zones 1, 2 et 3

Les autres modèles réalisés sont : ARIMA, NNAR, STL, TBATS et la combinaison de ces modèles.

Zone géographique 1



L'erreur quadratique moyenne de l'échantillon test est donnée dans le tableau suivant :

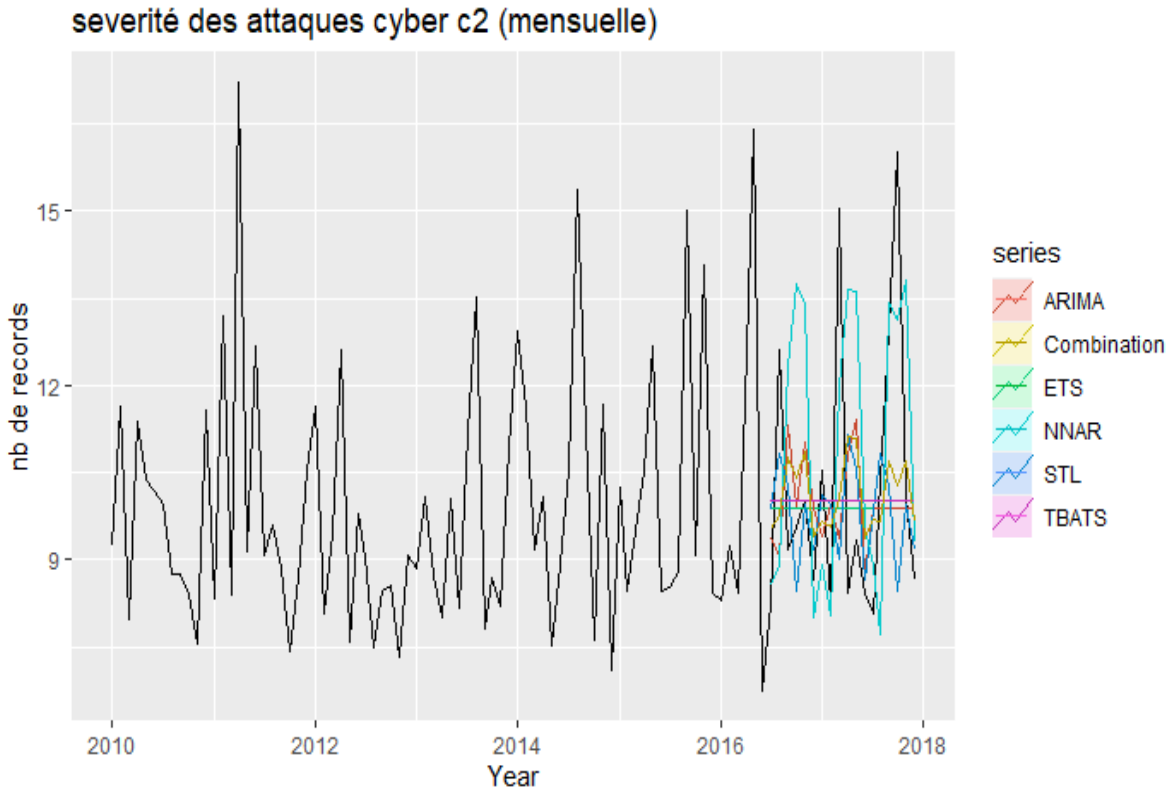
Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA (3)
RMSE	1.97	2.31	2.83	3.12	2.77	2.76	2.22

Figure 41 : Prédictions de la sévérité (performances) - Zone 1

C'est le modèle ets (A, N, N) qui propose l'erreur moyenne quadratique la plus petite, donc c'est le modèle le plus prédictif.

Toutefois, comme nous pouvons l'observer sur le graphique au-dessus, la prédiction reste très moyenne.

Zone géographique 2

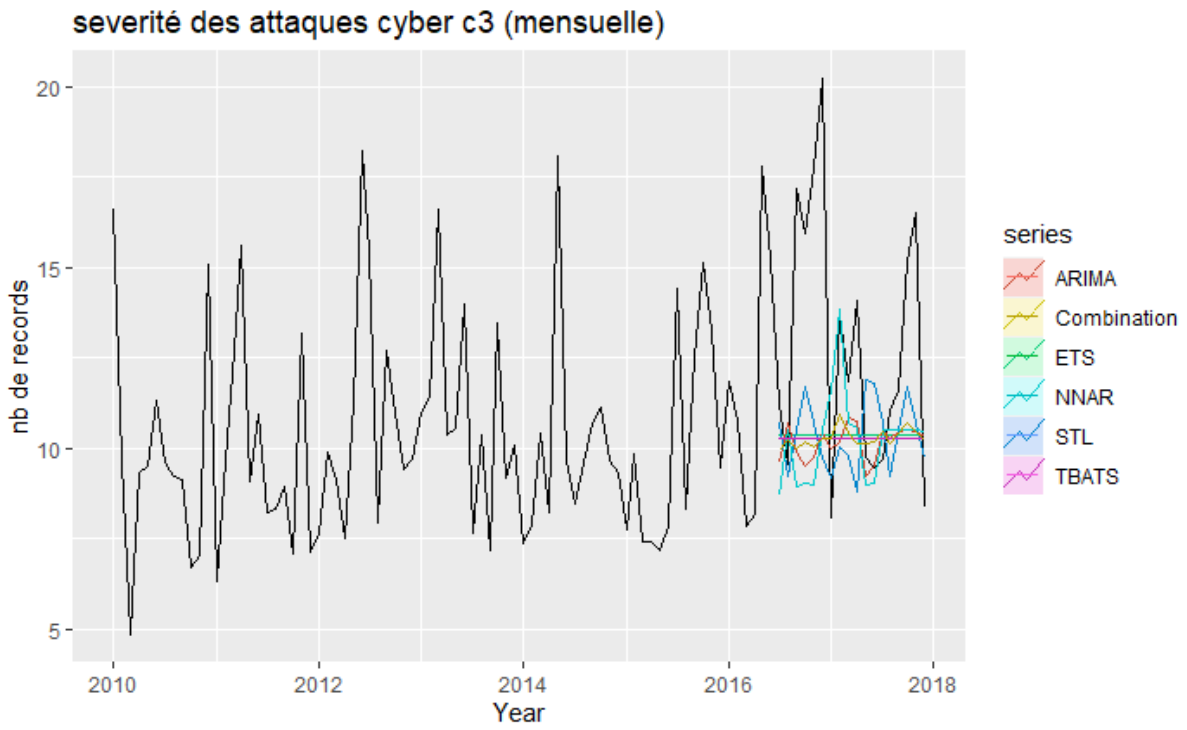


Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA
RMSE	2.31	2.14	2.47	2.48	2.31	2.25	2.39

Figure 42 : Prédictions de la sévérité (performances) - Zone 2

Le modèle qui est le plus prédictif pour la zone géographique 2 est le modèle ARIMA(0,0,0)(0,0,1)[12]

Zone géographique 3



Comme pour la zone géographique 1, le modèle le plus prédictif reste le modèle ETS(A,N,N).

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA(3)
RMSE	2.01	3.44	4.30	4.63	4.34	4.34	3.33

Figure 43 : Prédictions de la sévérité (performances) - Zone 3

Pour les trois zones géographiques, les résultats semblent très moyens.

La série temporelle étudiée sans distinguer les zones géographiques :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination	MA (3)
RMSE	2.52	2.47	2.55	3.15	2.66	2.76	2.49

En comparant les résultats obtenus par zone géographique avec une série initiale (sans distinction des zones), nous observons que la prédiction est améliorée pour les 3 zones.

- Fréquence de la base sans distinction de zone géographique

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	18.62	12.89	21.53	15.77	19.73	17.50

Nous pouvons également nous intéresser à la prédiction du nombre d'incidents mensuel pour les 3 zones géographiques.

Zone géographique 1 (nombre de sinistres par mois) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	9.90	11.32	12.46	6.22	10.16	9.80

Zone géographique 2 (nombre de sinistres par mois) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	3.79	4.95	3.62	6.61	3.48	4.11

Zone géographique 3 (nombre de sinistres par mois) :

Modèle	ETS	ARIMA	STL-ETS	NNAR	TBATS	Combination
RMSE	2.14	2.18	2.33	2.49	2.08	2.18

Figure 44 : Prédiction de la fréquence - Zone 1, 2 et 3

Nous observons que par rapport à la série qui ne distingue pas les zones géographiques, la performance est améliorée pour toutes les zones. Tous les modèles améliorent leurs performances.

C'est le modèle NNAR qui est le plus performant pour la zone géographique 1, TBATS pour les zones 2 et 3.

Les modèles de séries temporelles donnent des résultats relativement satisfaisants.

On peut voir la qualité de la prédiction et comparaison des modèles dans les tableaux récapitulatifs ci-dessus.

Le regroupement par zone géographique a eu un effet bénéfique sur les performances prédictives des modèles des séries temporelles.

On peut remarquer que les modèles de séries temporelles seraient moins performants que les modèles de Machine Learning pour la prédiction de la gravité d'incidents, alors que pour la fréquence, ils font mieux pour les zones géographiques 2 et 3.

Partie III - L'étude des dépendances observées entre les sinistres et leur modélisation par les méthodes de copules

L'étude de la dépendance est une étude du lien qui peut exister entre les variables. La dépendance et la corrélation sont des notions bien distinctes.

$X \perp\!\!\!\perp Y \rightarrow r(X,Y) = 0$ donc X et Y ne sont pas corrélés, mais la réciproque est fausse (sauf pour les cas où les variables sont gaussiennes, dans ce cas-là, c'est le coefficient de corrélation qui caractérise la dépendance)

La dépendance peut être mesurée de plusieurs manières :

- Le taux de corrélation de Pearson.
- Tau de Kendall
- Rho de Spearman

3.1 La mesure de la dépendance

3.1.1 Le coefficient de corrélation de Pearson

La manière la plus simple et la plus courante pour estimer la corrélation linéaire entre deux variables est le coefficient de Pearson. La formule pour calculer le coefficient de Pearson est la suivante :

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Ce coefficient $r \in [-1,1]$, si $r = 0$, il n'y a pas de corrélation entre les variables donc les variables sont indépendantes, une valeur négative de r signifie une corrélation négative, alors qu'une valeur positive montre une corrélation positive, c'est-à-dire que les deux variables varient dans le même sens.

Mais ce coefficient présente quelques limites. Il ne détecte pas les valeurs aberrantes dans les données et ne peut pas détecter les relations qui ne sont pas linéaires.

Par conséquent, r = 0 ne signifie pas forcément l'indépendance, mais simplement l'indépendance linéaire.

Le coefficient de Spearman est approprié lors de l'étude des données qui sont des distributions gaussiennes. Mais dans le domaine de l'assurance, les distributions respectent rarement ces conditions.

3.1.2 Le coefficient de rang de Spearman

Le coefficient de rang de Spearman permet de mesurer la corrélation pour le nombre de données faible, dans le cas où le coefficient de Spearman n'est pas adapté. Dans le cas de taux de rang de Spearman, les valeurs des variables sont remplacées par leur rang. Il est défini par la formule suivante :

$$\rho = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n^3 - n}$$

Où d_i représente une différence entre les rangs de X et Y, n, le nombre de ces couples (X, Y).

Cette formule est appliquée si l'hypothèse de normalité n'est pas respectée.

3.1.3 Le tau de Kendall

Le tau de Kendall (τ) est une mesure de corrélation de rang, il a été introduit par Kendall en 1938.

A la différence du coefficient de corrélation de Pearson, qui est basé sur l'hypothèse de linéarité entre les variables, le tau de Kendall serait une mesure d'association.

Soit(X_1, Y_1) et (X_2, Y_2) deux vecteurs iid (indépendants et de même loi). Alors on définit le tau de Kendall (τ) comme la probabilité de concordance moins la probabilité de discordance :

$$\tau = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

$$= P[(X_1 - X_2)(Y_1 - Y_2) > 0] - (1 - P[(X_1 - X_2)(Y_1 - Y_2) > 0])$$

$$= 2P[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1$$

La paire est concordante si (X_1, Y_1) et (X_2, Y_2) suivent la propriété suivante :

- $X_1 > X_2$ et $Y_1 > Y_2$ou
- $X_1 < X_2$ et $Y_1 < Y_2$

La paire est discordante si (X_1, Y_1) et (X_2, Y_2) suivent la propriété suivante :

- $X_1 > X_2$ et $Y_1 < Y_2$ou
- $X_1 < X_2$ et $Y_1 > Y_2$

Le tau de Kendall (τ) est invariant par transformation croissante qu'elle soit linéaire ou non, c'est le rang de chaque observation qui compte. La corrélation de Kendall entre deux variables sera d'autant plus élevée que les observations ont un rang similaire.

Si les deux rangs sont indépendants, τ serait égal à 0, dans le cas de rangs identiques, la valeur du τ serait égale à 1 et -1 en cas de désassociation entre les rangs.

Ce coefficient peut être défini par :

$$\tau_n = \frac{P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]}{\binom{n}{2}} - 1 = \frac{4C_n}{n(n-1)} - 1$$

Où $\binom{n}{2} = \frac{n(n-1)}{2}$ est le coefficient binomial du nombre de manières de choisir deux éléments parmi n éléments et donc C_n est le nombre de paires concordantes parmi les $\frac{n(n-1)}{2}$ paires de couples (X_i, Y_i) et $(X_j, Y_j$ lorsque $1 \leq i < j \leq n$.

La dépendance entre les variables peut être modélisée par ces trois indicateurs présentés, mais pour modéliser cette dépendance, une autre méthode est utilisée : les copules.

3.2 Les copules

La notion de copule a été utilisée pour la première fois par Sklar en 1959.

Pour modéliser la dépendance conjointe entre plusieurs variables, les copules sont l'outil le plus utilisé.

Cette sous-partie présentera un petit rappel de la théorie des copules : la définition, les propriétés, les caractéristiques principales et les différentes structures de copules.

Cette méthode a plusieurs avantages : elle permet de donner une représentation de la dépendance entre les risques mais aussi de décrire un comportement de chaque risque (loi marginale).

Une copule est une fonction de répartition multidimensionnelle ayant des lois marginales uniformes sur [0,1].

Elle est définie par $C(u, v) = P(U \leq u, V \leq v)$ de $[0,1]^2$vers [0,1] et qui vérifie les trois propriétés suivantes :

1. $C(u, 0) = C(0, u) = 0$;
2. $C(u, 1) = u$ Et $C(1, v) = v$;
3. Pour tout $u_1 \leq u_2$ *et* $v_1 \leq v_2$,

$$C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0$$

Le théorème de Sklar fait un lien entre la notion de copule et la fonction de répartition multivariée.

La copule peut être vue comme une fonction de répartition multivariée des variables aléatoires $F_X(x)$ et $F_Y(y)$.

Donc le théorème de Sklar permet de modéliser par une seule copule la dépendance entre plusieurs variables.

Théorème de Sklar : Soit F une fonction de distribution bidimensionnelle dont les marginales sont F1 et F2. Alors F admet une représentation copule suivante :

$$F(x,y) = C\big(F_X(x), F_Y(y)\big)$$

Si les lois marginales sont des lois continues, alors cette copule est unique.

Corollaire du théorème de Sklar :

Pour tout vecteur $(u,v) \in [0,1]^2$, nous avons :

$$C(u,v) = F\left(F_X^{-1}(u), F_Y^{-1}(v)\right)$$

La densité d'une copule c peut être définie par une formule suivante :

$$c(u,v) = \frac{\partial C(u,v)}{\partial u \partial v}$$

3.2.1 Les copules usuelles

- La **copule de survie** $\overline{C}$est définie de $[0,1]^2$ vers [0,1] par la formule suivante :

$$\overline{C}(u,v) = u + v - 1 + C(1-u, 1-v)$$

On peut démontrer très facilement cette formule.

La fonction de survie conjointe de X et Y est définie par :

$$\begin{gathered}\overline{F}_{XY}(x,y) = P(X > x, Y > y) \\ = P(X > x) - P(X > x, Y < y) \\ = P(X > x) - P(Y < y) + P(X < x, Y < y) \\ = 1 - F_X(x) - F_Y(y) + F_{XY}(x,y) \\ = S_X(x) + S_Y(y) - 1 + C\big(1 - S_X(x), 1 - S_Y(y)\big) \\ = u + v - 1 + C(1-u, 1-v)\end{gathered}$$

- La **copule d'indépendance**

La copule d'indépendance ou le copule produit est définie par :

$$\prod(X_1, X_2) = X_1 \cdot X_2$$

- Les **copules maximales et minimales**

La copule minimale ou la copule anti monotone est définie par la formule suivante :

$$C^-(X_1, X_2) = \max(0, X_1 + X_2 - 1)$$

Alors que la copule maximale (ou la copule comonotone) est définie par :

$$C^+(X_1, X_2) = \min(X_1, X_2)$$

Ces deux définitions nous permettent d'avoir les propriétés immédiates d'une copule bivariée à l'aide du théorème Bornes de Fréchet-Hoeffding.

Théorème (**Bornes de Fréchet-Hoeffding**).

Soit H une fonction de répartition conjointe d'un vecteur aléatoire (X_1, X_2) de fonctions de répartition marginales $F1$ et $F2$. Pour toute copule C associée à H, on a $\forall u_1, u_2 \in I$,

$C^-(u_1, u_2) \leq C(u_1, u_2) \leq C^+(u_1, u_2)$

La conséquence du théorème des Bornes de Fréchet est :

$\forall u_1, u_2 \in \mathrm{I}, \quad C^-(u_1, u_2) \leq \prod(u_1, u_2) \leq C^+(u_1, u_2)$

- **Les copules elliptiques**

La copule est elliptique si elle est définie à partir d'une famille des lois elliptiques. Une variable $X = (X_1, X_2)$ suit une loi elliptique si elle est caractérisée par la propriété suivante :

$$X = \mu + Z\Sigma^{1/2}U$$

Où μest la moyenne, Z est une v.a. > 0, U une v.a. uniformément distribuée sur le disque unitaire R^2et $Z \perp\!\!\!\perp U$

Les copules les plus connues qui font partie de cette famille sont : la copule gaussienne et la copule de Student.

- **La copule normale**

Cette copule suit une distribution normale bivariée, elle est définie par la formule suivante :

$$C_p^{Ga}(u, v) = \int_{\infty}^{\phi^{-1}(u)} \int_{\infty}^{\phi^{-1}(v)} \frac{1}{2\pi\sqrt{(1-\rho^2)}} exp\{\frac{t_1^2 + t_2^2 - 2\rho t_1 t_2}{2(\rho^2 - 1)}\} dt_1 dt_2$$

Avec, $-1 < \rho < 1$ et ϕ la fonction de distribution de la loi normale

La copule normale ne présente pas de caractère de queue épaisse, elle n'est pas vraiment adaptée à l'analyse des dépendances extrêmes.

- **La copule de Student**

La copule de Student est une copule paramétrée par deux coefficients ρ (coefficient de corrélation) et ν (degré de liberté), elle est associée à une distribution multivariée de Student :

$$C_{\rho,\nu}(u, v) = \int_{-\infty}^{t_\nu^{-1}(u)} \int_{-\infty}^{t_\nu^{-1}(v)} \frac{\Gamma\left(\frac{\nu+2}{2}\right)}{2\pi\Gamma\left(\frac{\nu}{2}\right)\sqrt{1-\rho^2}} \times \left(1 + \frac{s^2 + t^2 - 2\rho st}{\nu(1-\rho^2)}\right)^{-\frac{\nu+2}{2}} dsdt$$

Où t_v est la fonction de répartition de la loi de Student à ν degrés de libertés.

i.e.

$$t_\nu(x) = \int_{-\infty}^{x} \frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\sqrt{\rho\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} dt$$

Si le degré de liberté ν → +∞, la copule de Student → copule gaussienne.

Avantages des copules elliptiques :

Les distributions auxquelles elles sont associées sont bien connues, donc leur utilisation est relativement simple ;

Inconvénients des copules elliptiques :

Les copules elliptiques sont des copules symétriques et donc leurs coefficients de dépendance supérieurs et inférieurs sont les mêmes. Par conséquent, Il n'est pas possible de réaliser des modèles fiables pour les données présentant des dépendances de queue.

Par ailleurs, si le nombre des variables est assez grand par rapport au nombre d'observations, l'ajustement de copules se complique, car il résulte en une matrice non inversible ou mal conditionnée.

- **Les copules archimédiennes**

Les copules archimédiennes sont générés à partir d'une fonction ϕ appelé le générateur de la copule.

Les copules archimédiennes se présentent sous la forme générale :

$$C_\varphi(u_1, \dots, u_n) = \varphi^{-1}\big(\varphi(u_1) + \cdots + \varphi(u_n)\big)$$

$\varphi: [0,1] \Rightarrow [0,\infty]$ est une fonction continue, convexe et strictement décroissante telle que $\varphi(1) = 0$.

Soit $\varphi(0) = +\infty$ *si* $\lim_{x\to\infty} \varphi(x) \Rightarrow \infty$ et $\varphi^{-1}(x) = 0$ si $x \geq \varphi(0)$

La fonction φ est appelé générateur archimédien de la copule C.

Le tableau récapitulatif présente les copules archimédiennes les plus connues :

Copula Type	Copule	Paramètres	Générateur
Clayton	$\left[\sum_{i=1}^{m} u_i^{-\theta} - m + 1\right]^{\frac{-1}{\theta}}$	$\theta > 0$	$\theta^{-1}\left(u^{-\theta} - 1\right)$
Franck	$\frac{1}{\theta} log\left[1 + \frac{\Pi_{i=1}^{m}[exp(-\theta u_i) - 1]}{(exp(-\theta) - 1]^{m-1}}\right]$	$\theta \epsilon \frac{(-\infty,\infty)}{\{0\}}$ pour $m = 2$ et $\theta > 0$ pour m > 3	$-log\left[\frac{exp(-\theta u) - 1}{exp(-\theta) - 1}\right]$

Gumbel	$\exp\{-\left[\sum_{i=1}^{m}(-\log u_i)^{\theta}\right]^{\frac{1}{\theta}}\}$	$\theta > 1$	$(-logu)^{\theta}$

Figure 45 : Copules archimédiennes

Ces copules archimédiennes présentent plusieurs avantages :

- Facilité d'utilisation ;
- Elles sont bien adaptées pour la modélisation des dépendances extrêmes ;
- Un grand choix de familles de copules ;
- Utiles pour modéliser des risques échangeables

- **Les copules extrêmes**

Définition : Une fonction C est une copule extrême s'il existe une copule telle que :

$$\Gamma\,(u_1^{\frac{1}{n}},\dots,u_d^{\frac{1}{n}})^n \xrightarrow[n\to\infty]{} C\,(u_1,\dots,u_d)$$

Pour tout $(u_1,\dots,u_d) \in [0,1]^d$, Γ est dans le max-domaine d'attraction de C.

Le tableau recensant des copules extrêmes :

Famille,modèle	Copule $C_\theta(u,v)$	Générateur $A_\theta(t)$
Indépendante	$C(u,v) = \Pi(u,v) = uv$	$A(t) = 1$
Gumbel$_1$ ou logistique	$C_\theta(u,v) = \exp\{-(\widetilde{u^\theta} + \widetilde{v^\theta})^{\frac{1}{\theta}}\}$	$A_\theta(t) = [t^\theta + (1-t^\theta)]^{\frac{1}{\theta}}$
Gumel$_2$	$C_\theta(u,v) = uv\exp\{\theta\frac{\widetilde{uv}}{\tilde{u}+\tilde{v}}\}$	$A_\theta(t) = t^2 - \theta t + 1$
Galambos	$C_\theta(u,v) = uv\exp\{-(\widetilde{u^{-\theta}} + \widetilde{v^{-\theta}})^{\frac{1}{\theta}}\}$	$A_\theta(t) = 1 - [t^{-\theta} + (1-t)^{-\theta}]^{\frac{-1}{\theta}}$
Husler-Reiss	$C_\theta(u,v) = \exp\{-\tilde{v}\phi\left[\frac{1}{\theta}+\frac{1}{2}\theta\log\left(\frac{\tilde{v}}{\tilde{u}}\right)\right] - \tilde{u}\phi\left[\frac{1}{\theta}+\frac{1}{2}\theta\log\left(\frac{\tilde{u}}{\tilde{v}}\right)\right]\}$ $A_\theta(t) = t\phi\left[\frac{1}{\theta}+\frac{1}{2}\theta\log\left(\frac{t}{1-t}\right)\right] + (1-t)\phi\left[\frac{1}{\theta}-\frac{1}{2}\log\left(\frac{t}{1-t}\right)\right]$	
Logistique de Joe	$C_{\theta,\delta}(u,u) = \exp\{-\left[\tilde{u}^\theta + \tilde{v}^\theta - (\tilde{u}^{-\theta\delta} + \tilde{v}^{-\theta\delta})^{-\frac{1}{\delta}}\right]^{\frac{1}{\theta}}\}, \delta > 0, \theta \geq 1$ $A_{\theta,\delta}(t) = [t^\theta + (1-t)^\theta) - (t^{-\theta\delta} + (1-t)^{-\theta\delta})^{-\frac{1}{\delta}}]^{\frac{1}{\theta}}, t > 0$	
Galambos	$C_{\theta,\delta}(u,v) = \exp\{-\left[\tilde{u}^\theta + \tilde{v}^\theta - (\tilde{u}^{-\theta\delta} + \tilde{v}^{-\theta\delta})^{-\frac{1}{\delta}}\right]^{\frac{1}{\theta}}\}, \delta > 0, \theta \geq 1$ $A_{\theta,\delta}(t) = 1 - [t^{-\theta} + (1-t)^{-\theta} - (t^{\theta\delta} + (1-t)^{\theta\delta})^{\frac{1}{\delta}}]^{-\frac{1}{\theta}}$	
Marshal-Olkin	$C_{\alpha,\beta}(u,v) = u^{1-\alpha}v^{1-\beta}\min(u^\alpha, v^\beta) = \begin{cases} uv^{1-\beta} \; si \; u^\alpha < v^\beta \\ u^{1-\alpha} \; si \; u^\alpha > v^\beta \end{cases}$ $A_{\alpha,\beta}(t) = \max\{1-\alpha t, 1-\beta(1-t)\}, \alpha \leq 1, \beta \geq 0$	
Tawn	$C_{\theta,\delta,\lambda} = uv\exp\{-(1-\delta) + (\theta-\delta)\tilde{u} + [(\theta\tilde{u})^\lambda + (\delta\tilde{v})^\lambda]^{\frac{1}{\lambda}}\}$ $A_{\{\theta,\beta,\lambda\}}(t) = (1-\delta) + (\delta-\theta)t + [(\theta t)^\lambda + (\delta(1-t)^\lambda]^{\frac{1}{\lambda}}$	

Figure 46 : Copules extrêmes

3.3 La sélection de la bonne copule pour chaque type de risque

Dans cette partie, la structure de dépendance de la gravité des pertes mensuelles moyennes de différents types d'incidents cyber sera analysée.

Nous avons trois types de dépendances possibles :

- Les pertes selon la zone géographique,
- Les pertes intersectorielles (secteur médical, financier, commerce ou administration publique),
- Les pertes par type de violation subie (piratage, perte d'un appareil électronique, divulgation par une personne interne à l'entreprise, perte des documents papier).

Comme dans le cas des séries temporelles, on a procédé à un regroupement des sinistres par zone de risque homogène. Pour construire les copules, nous devons d'abord trouver les lois marginales des distributions pour chaque type de risque.

Dans l'ajustement de la sévérité, nous ne considérons que les valeurs non nulles. Plusieurs distributions continues sont testées : log-normale, log logistique, Pareto, exponentielle, normale, weibull et gamma.

La distribution la mieux adaptée aux données est évaluée en minimisant l'AIC et en effectuant le test de Kolmogorov-Smirnov (test K-S).

Nous pouvons aussi mesurer l'adéquation de la distribution aux données en comparant la log-vraisemblance. Cependant ce critère croit avec l'augmentation du nombre de paramètres. C'est pourquoi l'utilisation des critères de vraisemblance pénalisés (AIC ou BIC) ont été préférés.

Pour commencer, nous allons nous intéresser à la structure de dépendance entre les 3 zones de risque.

La modélisation de la dépendance entre X et Y sera effectuée par étapes :

1. La détermination des lois marginales et la vérification du bon ajustement de ces lois en minimisant AIC et BIC et en effectuant le test de Kolmogorov-Smirnov.
2. Identification de la copule à utiliser et l'estimation de ces paramètres.
3. On vérifie que la distribution est bien ajustée et on retient la copule qui minimise les critères d'Akaike (AIC) ou de Schwarz (BIC).

3.3.1 Modélisation de la dépendance entre les zones géographiques.

Le tableau suivant montre que la distribution Pareto est la mieux ajustée pour la plupart des distributions de la gravité des incidents cyber.

	Distributions	Paramètres	AIC	BIC	KS statistic
cluster 1	lnorm		2426,90	2432,03	0,14
	llogis	shape scale	2420,86	2425,99	0,10
	pareto	0,71 18296	**2416,38**	**2421,51**	0,10
	exp		2794,22	2796,78	0,65
	norm		3212,51	3217,64	0,43
	weibull		2474,86	2479,99	0,20
	gamma		2742,31	2747,44	0,77
cluster 2	lnorm		2340,65	2345,78	0,16
	llogis	shape scale	2335,17	2340,29	0,11
	pareto	0,56 6119,70	**2323,47**	**2328,60**	0,11
	exp		2784,90	2787,46	0,69
	norm		3165,41	3170,54	0,44
	weibull		2386,20	2391,33	0,19
	gamma		2572,55	2577,68	0,66
cluster 3	lnorm		2583,83	2588,96	0,15
	llogis	shape scale	2584,44	2589,57	0,11
	pareto	0,33 3671,86	**2566,60**	**2571,73**	0,08
	exp		3293,12	3295,68	0,73
	norm		3721,60	3726,72	0,43
	weibull		2617,37	2622,50	0,17
	gamma		2827,10	2832,23	0,69

Figure 47 *:* Ajustement des lois marginales en fonction des zones géographiques

Nous avons donc trouvé les distributions des lois marginales et de la copule qui modélise la structure de dépendance entre les zones géographiques.

Nous pouvons estimer la distribution jointe de la gravité des incidents cyber en fonction de type de violation et simuler des réalisations de cette distribution.

Les deux copules qui décrivent le mieux les structures de dépendance et sont donc retenues pour chaque cas de dépendance entre les zones géographiques sont les suivantes.

La copule de Gumbel est ajustée pour chaque type de dépendance, car cette copule fait partie des copules archimédiennes, mais c'est aussi une copule extrême. Elle ne capture que des dépendances positives, elle représente bien les structures de dépendance sur la queue supérieure. C'est pourquoi ce type de copule est adapté pour représenter des événements de forte intensité.

Par exemple, le lien entre la zone 1 et la zone 2 peut être modélisée par la copule de Gumbel (paramètre 1.0934) ou la copule de Joe (paramètres : 1.17 et 0.09).

Dépendances	Copules	paramètres	Copules gumbel paramètres
C2_C3	Rotated Tawn type 1 180 degrees	par = 3.7, par2 = 0.07, tau = 0.07	statistic = 0.075278, parameter = 1.0512, p-value =0.1274
C1_C2	Joe	par = 1.17, tau = 0.09	statistic = 0.027303, parameter = 1.0934, p-value =0.6169
C1_C3	Tawn type 2	par = 20, par2 = 0.03, tau = 0.03	statistic = 0.021066, parameter = 1, p-value = 0.8556

Figure 48 : Ajustement des copules en fonction des zones géographiques

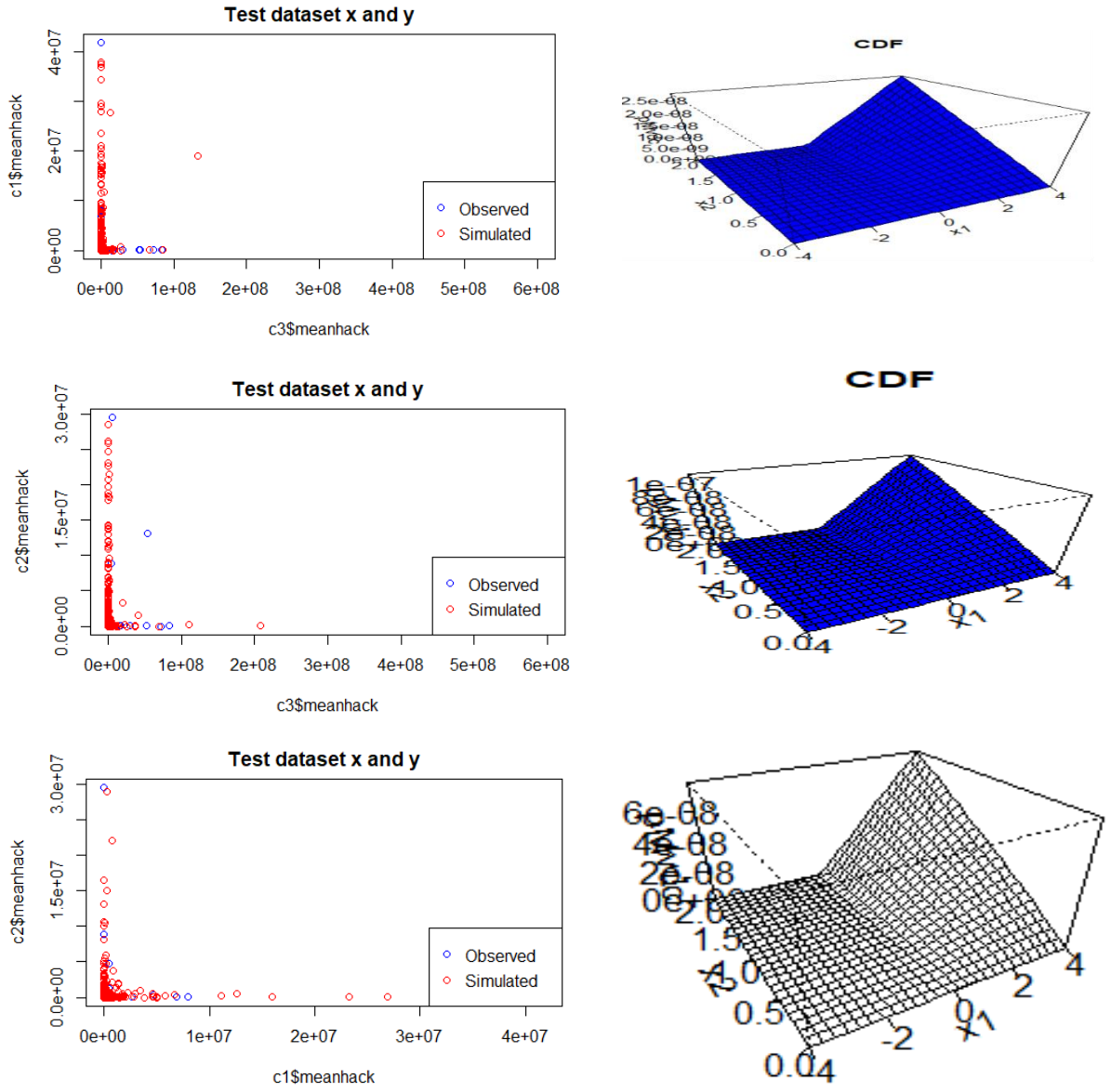


Figure 49 : Simulation des copules et CDF : Zone 1, 2 et 3

Les nuages de points de données réelles et des données simulées se superposent assez bien, l'adaptation de la copule aux données est donc suffisante.

Nous avons également fait un ajustement de copule spécifique pour chaque zone géographique en estimant les distributions des pertes mensuelles en fonction de type de violation **(int**, **hack, phys ou ord**) ou de secteur d'activité de l'entreprise.

Nous avons 6 combinaisons possibles entre différents types de violation pour chaque zone géographique, de même que pour les secteurs d'activité. Nous allons donc estimer 6 copules pour chaque dépendance sectorielle et 6 copules pour les dépendances entre les différents types de violation des données subiees pour chaque zone géographique.

En tout, on aura donc 36 copules à estimer.

3.3.2 Modélisation des dépendances entre les différents types de violation des données dans les 3 zones géographiques

- Zone géographique 1 (violation des données)

	Distributions	Paramètres	AIC	BIC
int	lnorm		1484,02	1488,55
	llogis	1,43 5939,64	**1479,35**	**1483,88**
	pareto		1486,72	1491,24
	exp		1599,35	1601,61
	norm		1889,56	1889,56
	weibull		1520,65	1525,17
	gamma		1732,66	1737,19
hack	lnorm		1838,56	1843,09
	llogis		1836,57	1841,09
	pareto	0,48 8585,12	**1835,23**	**1839,76**
	exp		2163,84	2166,10
	norm		2444,68	2449,20
	weibull		1857,31	1861,84
	gamma		1985,21	1989,74
phys	lnorm		1512,45	1516,97
	llogis	1,47 6177,40	**1505,81**	**1510,33**
	pareto		1517,15	1521,68
	exp		1606,03	1608,29
	norm		1825,71	1830,23
	weibull		1548,87	1553,40
	gamma		1651,41	1655,93
ord	**lnorm**	9,21 1,61	**1581,47**	**1585,99**
	llogis		1584,14	1588,67
	pareto		1584,59	1589,12
	exp		1643,87	1646,13
	norm		1801,88	1806,41
	weibull		1598,28	1602,81
	gamma		1626,18	1630,71

Figure 50 : Lois marginales des distributions pour chaque type de violation de données (zone 1)

Pour l'estimation de la gravité des incidents cyber touchant des états faisant partie de la zone 1, la loi log-logistique présente le meilleur ajustement pour les incidents cyber causés par les violations de type : **INT** et **Phys,** la loi de Pareto est mieux adaptée pour les violations dues au piratage **(HACK),** et la loi log-normale pour les incidents du type **ORD.**

Après avoir fait l'ajustement des distributions, nous pouvons identifier les copules à utiliser pour exprimer la structure de la dépendance entre les différents types de violations des données entre elles.

Dépendances	Copules	paramètres	Copules gumbel paramètres
hack_int	Independente		statistic = 0.047466, parameter = 1.0382, p-value =0.4281
hack_ord	Independente		statistic = 0.037602, parameter = 1.0127, p-value =0.5969
hack_phys	Independente		statistic = 0.032521, parameter = 1.0266, p-value=0.6638
ord_int	independente		statistic = 0.04071, parameter = 1.0679, p-value =0.4371
ord_phys	Frank	par = 1.07, tau = 0.12	statistic = 0.038792, parameter = 1.0926, p-value =0?4201
int_phys	Rotated Tawn type 2 90 degrees	par = -5.67, par2 = 0.06, tau = -0.06	statistic = 0.056387, parameter = 1, p-value = 0.3571

Figure 51 : Identification des copules (zone 1)

- Zone géographique 2 (violation des données)

Pour l'estimation de la gravité des incidents cyber touchant des états faisant partie de la zone 2, la loi Pareto présente le meilleur ajustement pour les incidents cyber causés par des violations de type : **INT**, **HACK** et **Ord,** la loi log-logistique est mieux adaptée pour les violations de type **Phys.** Les coefficients des distributions qui donnent le meilleur ajustement sont donnés dans le tableau suivant.

	Distributions	Paramètres		AIC	BIC
int	lnorm			1155,98	1159,81
	llogis			1152,36	1156,19
	pareto	0,70	5846,30	**1150,11**	**1153,94**
	exp			1338,68	1340,59
	norm			1542,37	1546,19
	weibull			1177,16	1180,99
	gamma			1290,70	1294,53
hack	lnorm			1193,17	1196,99
	llogis			1266,66	1270,49
	pareto	0,40	1782,39	**1192,88**	**1196,70**
	exp			1523,23	1525,15
	norm			1739,08	1742,90
	weibull			1209,40	1213,23
	gamma			1317,21	1321,03
phys	lnorm			1000,79	1004,61
	llogis	1,30	2951,62	**994,12**	**997,95**
	pareto			998,30	1002,13
	exp			1131,14	1133,06
	norm			1332,22	1336,04
	weibull			1029,48	1033,30
	gamma			1156,85	1160,68
ord	lnorm			1048,94	1052,76
	llogis			1045,91	1049,74
	pareto	0,77	2597,96	**1044,73**	**1048,55**
	exp			1193,52	1195,43
	norm			1371,43	1375,25
	weibull			1071,81	1075,64
	gamma			1152,87	1156,69

Figure 52 : Lois marginales des distributions pour chaque type de violation de données (zone 2)

Comme précédemment, on peut identifier les copules à utiliser pour exprimer la structure de la dépendance entre les différents types de violations des données entre elles.

Dépendances	Copules	paramètres	Copules gumbel paramètres		
hack_int	**Independente**		statistic = 0.045665,	parameter = 1.0878,	p-value = 0.3761
hack_ord	**Independente**		statistic = 0.080064,	parameter = 1.0384,	p-value = 0.1434
hack_phys	**Rotated Tawn type 2 90 degrees**	par = -3.65, par2 = 0.12, tau = -0.11	statistic = 0.0036386,	parameter = 1,	p-value = 0.9965
ord_int	**independente**		statistic = 0.065415,	parameter = 1,	p-value = 0.2792
ord_phys	**Rotated Tawn type 2 180 degrees**	par = 3.35, par2 = 0.14, tau = 0.13	statistic = 0.021126,	parameter = 1.1476,	p-value = 0.6668
int_phys	**Independente**		statistic = 0.040371,	parameter = 1,	p-value = 0.6049

Figure 53 : Identification des copules (zone 2)

- Zone géographique 3 (violation des données)

Pour l'estimation de la gravité des incidents cyber touchant des états faisant partie de la zone 3, la loi log-normale présente le meilleur ajustement pour les incidents cyber causés par les violations de type : **INT**, **HACK** et **Phys,** la loi Pareto est mieux adaptée pour les violations de type **Ord.**

	Distributions	Paramètres		AIC	BIC
int	**lnorm**	8,92	1,95	**1016,58**	**1020,24**
	llogis			1018,76	1022,42
	pareto			1019,88	1023,53
	exp			1050,30	1052,13
	norm			1156,05	1159,70
	weibull			1017,90	1021,56
	gamma			1032,87	1036,53
hack	**lnorm**	10,32	3,86	**1208,75**	**1212,40**
	llogis			1210,78	1214,44
	pareto			1210,67	1214,33
	exp			1543,08	1544,91
	norm			1701,56	1705,21
	weibull			1218,33	1221,98
	gamma			1249,74	1253,40
phys	**lnorm**	8,77	2,01	**967,50**	**971,15**
	llogis			969,87	973,53
	pareto			970,05	973,71
	exp			1011,18	1013,01
	norm			1131,93	1135,58
	weibull			978,71	982,37
	gamma			1012,88	1016,54
ord	lnorm			1005,66	1009,32
	llogis			1004,50	1008,16
	pareto	1,02	5081,13	**1002,74**	**1006,39**
	exp			1127,22	1129,05
	norm			1276,06	1279,72
	weibull			1021,67	1025,33
	gamma			1073,71	1077,36

Figure 54 : Lois marginales des distributions pour chaque type de violation de données (zone 3)

Dépendances	Copules	paramètres	Copules gumbel paramètres		
hack_int	**Rotated Tawn type 2 180 degrees**	par = 20, par2 = 0.07, tau = 0.07	statistic = 0.085593,	parameter = 1.0252,	p-value = 0.1533
hack_ord	**Rotated Clayton 270 degrees**	par = -0.58, tau = -0.23	statistic = 0.09565,	parameter = 1,	p-value = 0.1064
hack_phys	**Rotated Tawn type 1 180 degrees**	par = 20, par2 = 0.05, tau = 0.05	statistic = 0.12385,	parameter = 1.0109,	p-value = 0.0225
ord_int	**Rotated Tawn type 1 90 degrees**	par = -8, par2 = 0.05, tau = -0.05	statistic = 0.085565,	parameter = 1.009,	p-value = 0.1364
ord_phys	**Rotated Tawn type 2 180 degrees**	par = 20, par2 = 0.07, tau = 0.07	statistic = 0.056335,	parameter = 1.073,	p-value = 0.2742
int_phys	**Independente**		statistic = 0.031075,	parameter = 1.080,	p-value = 0.6279

Figure 55 : Identification des copules (zone 3)

3.3.3 Modélisation des dépendances sectorielles dans les différentes zones géographiques.

- Zone géographique 1 (secteur d'activité)

	Distributions	Paramètres		AIC	BIC
med	lnorm			999,55	1003,21
	llogis	1,47	7493,83	**998,38**	**1002,03**
	pareto			1005,41	1009,07
	exp			1034,16	1035,99
	norm			1140,75	1144,41
	weibull			1018,66	1022,31
	gamma			1047,93	1051,59
adm	lnorm			1077,46	1081,12
	llogis	0,94	15512,84	**1077,18**	**1080,83**
	pareto			1077,37	1081,02
	exp			1120,89	1122,72
	norm			1244,88	1248,54
	weibull			1081,26	1084,92
	gamma			1113,90	1117,56
retail	lnorm			1004,02	1007,68
	llogis			1003,99	1007,64
	pareto	0,36	405,63	**1001,70**	**1005,35**
	exp			1298,35	1300,18
	norm			1472,26	1475,92
	weibull			1019,77	1023,43
	gamma			1080,85	1084,50
bsf	**lnorm**	9,58	3,64	**1135,08**	**1138,74**
	llogis			1135,64	1139,30
	pareto			1136,25	1139,90
	exp			1519,20	1521,03
	norm			1697,78	1701,44
	weibull			1147,60	1151,26
	gamma			1204,60	1208,26

Figure 56 : Lois marginales des distributions pour chaque secteur (zone 1)

Dans la zone géographique 1, la loi log-logistique présente le meilleur ajustement pour les incidents cybers causés à des entreprises des secteurs : **Med** et **Adm,** alors que la loi de paréto est mieux adaptée au secteur de retail (commerce) et la loi log-normale pour les incidents qui touchent le secteur BSF**.**

Après avoir fait l'ajustement des distributions, nous pouvons identifier les copules à utiliser pour exprimer la structure de la dépendance entre les différents secteurs.

Dépendances	Copules	paramètres	Copules gumbel paramètres
med_adm	Independente		statistic = 0,073, parameter = 1, p-value =0,246
med_retail	Rotated Tawn type 2 270 degrees	par = -3.02, par2 = 0.37, tau = -0.3	statistic = 0.028, parameter = 1, p-value = 0.774
med_bsf	Independente		statistic = 0.056, parameter = 1, p-value = 0.370
adm_retail	Bivariate copula: Joe	par = 1.35, tau = 0.17	statistic = 0.014, parameter = 1.214, p-value = 0.695
adm_bsf	Independente		statistic=0.064, parameter = 1, p-value = 0.323
retail_bsf	Rotated Tawn type 2 180 degrees	par = 6.39, par2 = 0.11, tau = 0.11	statistic = 0.045, parameter = 1.205, p-value = 0.194

Figure 57 : Identification des copules (zone 1)

- Zone géographique 2 (secteur d'activité)

Dans la zone géographique 2, la loi log normale présente le meilleur ajustement pour les incidents cyber causés aux entreprises des secteurs public (Adm), financier (BSF) et commercial (retail)**,** la loi pareto est mieux adaptée pour les violations de type **Med.** Les coefficients des distributions qui donnent le meilleur ajustement sont donnés dans le tableau suivant.

	Distributions	Paramètres		AIC	BIC
med	lnorm			673,37	676,24
	llogis			673,22	676,08
	pareto	0,74	3355,73	**672,07**	**674,94**
	exp			725,11	726,54
	norm			802,85	805,71
	weibull			684,28	687,14
	gamma			699,10	701,96
adm	**lnorm**	8,78	2,63	**696,50**	**699,37**
	llogis			697,06	699,92
	pareto			700,00	702,86
	exp			791,96	793,39
	norm			887,05	889,92
	weibull			701,27	704,14
	gamma			722,97	725,84
retail	**lnorm**	8,43	3,58	**693,79**	**696,66**
	llogis			694,03	696,90
	pareto			694,54	697,41
	exp			1028,81	1030,24
	norm			1161,54	1164,41
	weibull			704,58	707,45
	gamma			759,35	762,22
bsf	**lnorm**	9,40	3,87	**758,41**	**761,28**
	llogis			760,25	763,12
	pareto			763,64	766,51
	exp			986,93	988,36
	norm			1109,71	1112,58
	weibull			763,11	765,97
	gamma			802,42	805,29

Figure 58 : Lois marginales des distributions pour chaque secteur (zone 2)

Après avoir fait l'ajustement des distributions, nous pouvons identifier les copules à utiliser pour exprimer la structure de la dépendance entre les différents secteurs.

Dépendances	Copules	paramètres	Copules gumbel paramètres
med_adm	Rotated Clayton 270 degrees	par = -0.59, tau = -0.23	statistic = 0.054, parameter = 1, p-value = 0.396
med_retail	Independente		statistic = 0.064, parameter = 1.063, p-value = 0.239
med_bsf	Independente		statistic = 0.018, parameter = 1, p-value = 0.873
adm_retail	Independente		statistic = 0.055, parameter = 1.094, p-value = 0.301
adm_bsf	Independente		statistic = 0.083, parameter = 1.039, p-value = 0.143
retail_bsf	Rotated Tawn type 1 180 degrees	par = 20, par2 = 0.07, tau = 0.07	statistic = 0.090, parameter = 1.022, p-value = 0.107

Figure 59 : Identification des copules (zone 2)

- Zone géographique 3 (secteur d'activité)

Dans la zone géographique 3, la loi pareto présente le meilleur ajustement pour les incidents cyber causés aux entreprises des secteurs médical (Med), commercial (retail) et financier (Bsf) : la loi weibull est mieux adaptée pour estimer la sinistralité des entreprises du secteur public (Adm).

	Distributions	Paramètres		AIC	BIC
med	lnorm			494,66	496,84
	llogis			493,15	495,33
	pareto	0;73	4472,63	**492,23**	**494,41**
	exp			552,60	553,69
	norm			621,75	623,93
	weibull			504,33	506,51
	gamma			526,30	528,48
adm	lnorm			439,16	441,34
	llogis			440,25	442,43
	pareto			443,94	446,13
	exp			478,31	479,40
	norm			541,41	543,59
	weibull	044	6428,27	**439,14**	**441,32**
	gamma			450,34	452,52
retail	lnorm			540,01	542,19
	llogis			540,28	542,46
	pareto	0,23	172,76	**533,90**	**536,08**
	exp			766,19	767,28
	norm			836,11	838,29
	weibull			547,47	549,65
	gamma			557,15	559,34
bsf	lnorm			414,65	416,83
	llogis			414,29	416,47
	pareto	0,30	41,43	**411,57**	**413,75**
	exp			637,48	638,57
	norm			724,42	726,60
	weibull			423,87	426,06
	gamma			453,04	455,22

Figure 60 : Lois marginales des distributions pour chaque secteur (zone 3)

Après avoir fait l'ajustement des distributions, on peut identifier les copules à utiliser pour exprimer la structure de la dépendance entre les différents secteurs.

Figure 61 : Identification des copules (zone 3)

Dépendances	Copules	paramètres	Copules gumbel paramètres
med_adm	Independente		statistic = 0.066, parameter = 1.069, p-value = 0.253
med_retail	Rotated Tawn type 1 180 degrees	par = -20, par2 = 0.1, tau =- 0.1	statistic = 0.052, parameter = 1, p-value = 0.444
med_bsf	Rotated Joe 270 degrees	par = -1.45, tau = -0.2	statistic = 0.020, parameter = 1, p-value = 0.821
adm_retail	Rotated Tawn type 2 270 degrees	par = -4.88, par2 = 0.28, tau = -0.26	statistic = 0.021, parameter = 1, p-value = 0.807
adm_bsf	Rotated Gumbel 270 degrees	par = -1.38, tau = -0.27	statistic = 0.044, parameter = 1, p-value = 0.495
retail_bsf	Independente		statistic = 0.022, parameter = 1.090, p-value = 0.7038

Partie IV - Application dans le cadre assurantiel

Au-delà des analyses ayant fait l'objet des parties 2 et 3 de ce mémoire, nous avons souhaité nous intéresser à un exemple d'application assurantielle du risque cyber.

Il nous a semblé intéressant de nous pencher sur un exemple de tarification d'une garantie contre la violation de données à destination d'entreprises.

Pour ce faire, nous avons entrepris la construction d'un portefeuille d'assurés par le biais de la base PRC, en répliquant dans un premier temps un portefeuille de sinistres fictif sur 10 années, de 2013 à 2022.

Sur la base de ce portefeuille fictif, nous avons appliqué les modèles les plus performants de notre Partie 2 dans le but d'estimer la sinistralité à venir en année N+1, soit en 2023.

Pour rappel, nous estimons plus spécifiquement le logarithme de cette variable. Nous espérons pouvoir en déduire assez précisément une estimation du nombre d'enregistrements à venir en année N+1.

Dans notre modèle coût-fréquence, la fréquence des sinistres est un paramètre d'étude, dont nous analyserons l'impact sur le tarif.

En effet, la table PRC étant une base de sinistres, il n'est pas possible d'en ressortir quelconque information sur la fréquence à laquelle les sinistres surviennent, il aurait fallu pour cela disposer d'un portefeuille d'assurés.

Quant au coût, nous avons fait le choix de nous baser sur des modèles d'assurance paramétrique, ceux mis au point par Jacobs puis Farkas.

Notre choix s'est orienté vers cette solution car la base PRC ne fournit aucune information sur le coût réel des incidents cyber qui y sont répertoriés. Toutefois, celle-ci restitue l'information qui fait l'objet d'une étude dans ce mémoire, à savoir le nombre d'enregistrements compromis (i.e. : nombre de données violées). Cette variable peut être transformée en coût via les modèles d'assurance paramétrique mentionnés précédemment.

L'objet de cette partie est d'analyser sur la base d'un portefeuille de violation de données inspiré de la table PRC, la sensibilité d'un tarif de prime pure d'assurance aux variables et facteurs suivants :

- Tarif segmenté vs tarif global
- Impact d'une franchise
- Impact d'un plafond d'indemnisation
- Sensibilité à la fréquence des sinistres

Nous nous attarderons donc dans un premier temps aux méthodes utilisées pour la création du portefeuille fictif étudié.

Enfin, nous reviendrons sur les principes de l'assurance paramétrique aboutissant aux modèles utilisés dans notre approche, les commenterons et justifierons leur pertinence.

Pour finir, nous présenterons nos résultats et conclurons sur les enseignements à tirer d'un point de vue économique des analyses de sensibilité menées.

4.1 Construction du portefeuille de sinistres

Dans le but de donner un aspect plus pratique à ce mémoire et d'appliquer les modèles développés, nous avons entrepris la construction d'un portefeuille d'assurés. Ce portefeuille simulé est basé sur l'historique de sinistres cyber issu de la base PRC.

Pour commencer, nous avons calibré des lois de probabilité du nombre d'incidents mensuels en fonction de la zone géographique et du secteur d'activité de l'entreprise.

Nous avons également procédé au calibrage des lois de probabilité du nombre d'enregistrements compromis par incident (gravité du risque) mais avons toutefois dû limiter la gravité des sinistres utilisés pour le calibrage des lois de distribution à 300 000 par incident afin de ne pas biaiser la calibration par des valeurs trop extrêmes.

Par ailleurs, la prise en compte de ces valeurs ne reflétant pas la réalité du risque, pourrait entraîner des conséquences néfastes sur les performances dans la prédiction des modèles utilisés. La calibration est faite par la méthode du maximum de vraisemblance. Le meilleur ajustement des distributions a été déterminé en minimisant le critère AIC.

Dans un deuxième temps, nous avons procédé à la simulation de 10 000 observations pour chaque répartition (fréquence mensuelle et la gravité des sinistres) pour chaque type de risque (secteur activité + zone géographique) en créant donc une base des données simulées pour la fréquence (nombre de sinistres) et la gravité des sinistres (nombre d'enregistrements compromis par sinistre) pour chaque type de risque.

Finalement, nous avons effectué un tirage aléatoire dans cette base de la fréquence mensuelle pour chaque type de risque et généré des incidents cyber dans notre portefeuille fictif.

Il contient 7923 observations sur une période de 2013 à 2022.

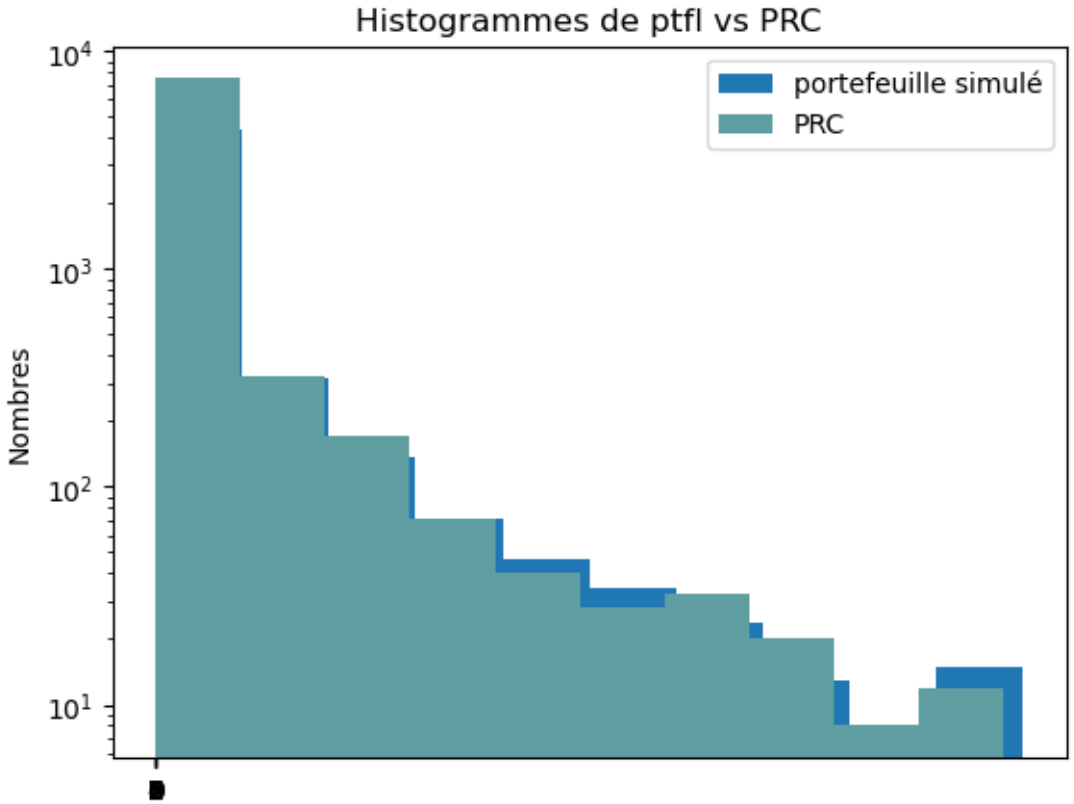


Figure 62 : Histogramme de portefeuille simulé vs PRC

Les histogrammes de variables de gravité de portefeuille simulé et de la base PRC se superposent relativement bien, démontrant que le portefeuille fictif reconstitué réplique convenablement la qualité et la profondeur de la table PRC.

Par la suite, nous appliquons les modèles de Machine Learning issus de la Partie 2 sur ce portefeuille fictif.

La métrique utilisée pour choisir le modèle le plus pertinent est la RMSE. Voici le tableau récapitulatif obtenu :

Modèle	NB d'incidents		Log(gravité)	
	Naif	Grid search	Naif	Grid search
Arbres de décision	6.503	4.778	1.431	1.376
Random Forest	5.688	4.501	1.433	1.422
Gradient Boosting	5.042	4.579	1.382	1.375
Réseaux de neurones	4.323	4.333	1.323	1.225

Figure 63 : Performances des modèles de ML sur le portefeuille simulé

Le modèle de réseaux de neurones est le plus performant parmi les modèles utilisés dans l'étude de la prédiction sur le portefeuille simulé. De la même manière que pour la base PRC, les performances sont plutôt satisfaisantes.

Nous rappelons que la modélisation a été effectuée sur la variable qui a été log transformée.

Par conséquent, même une bonne prédiction peut rapidement se dégrader du fait de la transformation inverse (exponentielle).

En estimant la sévérité de l'année 2023, nous constatons que le nombre d'enregistrements total prédit par les modèles n'est que de 11 millions alors que la moyenne annuelle d'enregistrements compromis sur le portefeuille simulé se situe aux environs 14 millions. Cela met en évidence l'effet combiné de l'hétérogénéité des données et de la transformation logarithmique de la gravité, qui est extrêmement sensible.

Nous estimons également que ce déphasage de la prédiction du log de la gravité avec le log cible d'environ 4% se traduit par un écart de la gravité de l'ordre de 20 à 25%.

Dès lors, il serait intéressant de pousser l'étude de la prédiction de la gravité (via la transformation logarithmique) et d'étudier la détermination d'un facteur d'ajustement du log, permettant une meilleure prédiction de la gravité.

Devant ces résultats, nous avons par conséquent choisi d'étudier les sensibilités du tarif sur un portefeuille 2023 simulé plutôt que sur le portefeuille estimé par les prédictions du modèle de la Partie 2.

4.2 Méthode de tarification

Le but de la modélisation est donc de bien modéliser la sinistralité afin de déterminer le tarif. L'étude de la tarification est limitée à la tarification de la prime pure qui représente l'espérance des pertes probables.

Cela signifie que notre analyse ne concernera pas la prime commerciale, étant notamment en manque d'information sur les chargements à appliquer.

Dans cette partie, nous estimerons le coût selon deux approches paramétriques (Jacobs, Farkas) basées sur l'étude de l'Institut de Ponemone du coût d'une violation de données.

Dans notre étude, nous utiliserons la tarification standard IARD : coût × fréquence. Cette méthode consiste à estimer le coût moyen d'un sinistre et sa fréquence.

La mutualisation doit être suffisante pour que cette approche soit véritablement viable. Ce n'est pas le cas pour le moment pour le risque cyber, comme on a vu dans l'introduction.

Nous n'avons pas de données fiables concernant la fréquence des sinistres cyber par secteur d'activité. Il est difficile de déterminer une fréquence précise du risque cyber pour chaque secteur d'activité.

Plusieurs facteurs peuvent l'impacter telle que la taille de l'entreprise, la nature de ses activités, les mesures de sécurité informatique mises en place, etc. Cependant, on peut se baser sur plusieurs études pour avoir une idée sur la fréquence.

Il est important de signaler que ces études n'utilisent que les incidents signalés, nous n'avons pas d'information sur l'ampleur des incidents non détectés ou non signalés par les victimes.

4.2.1 Coût

Les formules de Jacobs et de Farkas utilisées pour estimer le coût d'une violation de données sont basées sur le rapport de l'institut de Ponemone.

Ce dernier est spécialisé dans la recherche sur la protection des données.

Il publie un rapport annuel sur le coût des violations de données basé sur une enquête auprès des entreprises.

Comme vu précédemment, la base PRC possède une variable Totale.Records qui indique le nombre d'enregistrements compromis.

Sur la base des travaux de l'Institut de Ponemone, J. Jacobs a proposé une formule pour estimer le coût d'une violation des données :

$$log(Cout) = 7.68 + 0.76 \times log(Nb\ of\ records)$$

Il est important de préciser que dans son étude n'étaient pris en compte que les violations dont le nombre d'enregistrements ne dépassaient pas 100 000 données.

C'est pourquoi Farkas a réestimé la formule pour prendre en compte ces méga-violations. La formule qu'il a proposée est la suivante :

$$log(Cout) = 9.59 + 0.57 \times log(Nb\ of\ records)$$

Nous pouvons estimer le nombre d'enregistrements qui égalise les deux formules à 23 217. Pour les valeurs inférieures à 23 217 l'estimation par la formule de Farkas donne un coût supérieur à celui estimé par la formule de Jacobs et réciproquement.

Nous pouvons également préciser que les modèles développés estiment la sinistralité mensuelle et non pas la sinistralité pour chaque violation.

Ce point est important car le coût marginal d'une donnée violée est dépendant du nombre de données violées.

Les deux formules proposées sont un exemple d'assurance paramétrique.

Nous n'estimons pas le coût réel d'une violation de données, mais en se basant sur un indice, dans notre cas le nombre d'enregistrements compromis, nous estimons le coût approximatif d'une violation de données.

4.2.1.1 Assurance paramétrique

L'assurance paramétrique ne s'est réellement développée que depuis les années 1990 et donc une forme de couverture très récente. L'objectif premier de ce type de contrat était de se protéger face aux risques des catastrophes naturelles par le biais de contrats inspirés des dérivées climatiques mais paramétrables (période, indice, …).

Contrairement à une assurance traditionnelle qui reverse des indemnités en fonction des dégâts réels constatés par les experts, l'assurance paramétrique se base essentiellement sur la mesure d'un indice spécifique pour établir le montant des indemnités. Cet indice doit être facilement mesurable. L'un des principaux avantages de ce type d'assurance est la rapidité des paiements. Si l'assurance classique est plutôt adaptée à des risques d'intensité et de fréquence relativement faibles comme l'assurance automobile, mais semble moins appropriée aux risques rares et extrêmes que constituent les risques cyber, mais aussi des risques climatiques. On peut également citer d'autres avantages de l'assurance paramétrique tels que l'absence d'asymétrie d'information. L'assureur et l'assuré disposent de la même information donnée par la mesure d'un paramètre et le fait que l'antisélection et l'aléa moral sont fortement réduits.

Généralités sur l'assurance paramétrique

Pour pouvoir mesurer la vulnérabilité d'un portefeuille face au risque cyber, on pourrait définir un indice de sinistralité, c'est-à-dire un élément permettant de mesurer l'intensité de l'incident cyber et donc des dommages subis par les entreprises.

Définition d'un indice de sinistralité cyber

Dans le cas du risque cyber, une couverture paramétrique face aux attaques cyber par le biais d'un indice lié au nombre d'enregistrements compromis par cette attaque semble la plus évidente et simple à mesurer.

L'indice paramétrique pourrait aussi être composé en tenant compte de plusieurs facteurs, non seulement Total.Records (nombre de personnes touchées par cet incident), mais aussi en considérant les variables telles que type de violation, taille de l'entreprise, son secteur d'activité, sa vulnérabilité (mesures de prévention contre des incidents éventuels : antivirus, mises à jour, Service IT, cloud, etc) et la région où l'entreprise est située.

Les pondérations liées au paramètre géographique sont censées refléter l'importance économique de la zone géographique concernée et traduisent leur exposition au risque cyber. De même, certaines entreprises sont plus visées que d'autres par les pirates, c'est le cas, par exemple, des entreprises du secteur médical qui détiennent des données sensibles qui sont tout particulièrement prisées sur le darkweb.

Dans l'idéal, cet indice doit refléter les dégâts réels causés par l'événement assuré. Par conséquent, il est important d'établir tout d'abord une fonction de dommage qui relie les pertes économiques réelles à des indicateurs paramétriques mesurés lors de divers événements.

4.2.2 Fréquence

Ne disposant pas de données spécifiques permettant la création d'un portefeuille d'assurés, il nous a fallu procéder à la simulation d'un portefeuille de sinistres sur la base de la table PRC.

Cette approche peut paraître simpliste, dans la mesure où la fréquence des sinistres ici simulée ne relève d'aucune étude particulière et fait l'objet d'un paramétrage simple.

Il nous a cependant semblé intéressant d'analyser la sensibilité du tarif mis en place à la fréquence des sinistres, compte tenu de l'évolutivité de cet aspect-là du risque cyber.

Ainsi, nous avons modélisé et appliqué différentes fréquences par type d'organisation et par année de survenance, afin de tenir compte de caractéristiques de risques différentes selon ces axes d'analyse, ainsi que de l'évolutivité du risque.

Il serait également possible de moduler le tarif en fonction de zone géographique.

L'outil mis en place permet ainsi de visualiser rapidement et facilement l'impact d'une aggravation et/ou amélioration du risque du point de vue de sa composante « fréquence ».

4.3 Analyse des sensibilités

L'objet de cette partie est d'étudier la rentabilité d'un portefeuille cyber fictif par le biais d'une analyse des sensibilités d'un tarif de prime aux variables et facteurs suivants :

- Tarif segmenté vs tarif global
- Impact d'une franchise
- Impact d'un plafond d'indemnisation
- Sensibilité à la fréquence des sinistres

Ne disposant de telles données, dans notre étude, nous n'avons pas pris en compte ni la taille de l'entreprise, ni son chiffre d'affaires.

Nous considérons qu'il s'agit d'un manque important, qui pourrait permettre d'apporter de nouvelles solutions en matière de tarification.

Il s'agirait d'un axe intéressant à étudier dans des études postérieures.

Plus précisément, il pourrait être intéressant de mettre en parallèle la taille des entreprises ou leur CA avec le nombre d'enregistrements compromis, ce qui permettrait potentiellement d'ajuster les tarifs d'assurance des entreprises à leur taille. En ce sens cette approche permettrait potentiellement de faire bénéficier de tarifs plus compétitifs aux TPE et PME.

Cet axe d'étude-là n'a pas été abordé à ce stade dans nos travaux en raison de l'absence d'information sur la taille des entreprises dans la base PRC.

Taille du portefeuille

La taille du portefeuille est un élément important dans la tarification : plus il y a d'assurés plus la mutualisation des risques devient possible.

Toutefois, l'augmentation du volume de portefeuille ne rime pas toujours avec la baisse des tarifs, car si le portefeuille est composé d'un seul type de risque, il devient vulnérable.

A l'origine, notre portefeuille simulé est un portefeuille de sinistrés.

Sensibilité à la fréquence des sinistres

Il est difficile de donner une probabilité de subir un incident cyber dans l'année. Plusieurs facteurs peuvent influer sur la survenance des sinistres.

Il nous a toutefois semblé intéressant d'analyser la sensibilité du tarif mis en place à différents jeux de fréquence de sinistre, compte tenu de l'évolutivité de cet aspect-là du risque cyber.

Certaines études permettent d'avoir une certaine idée sur la situation réelle : Vérizon, Hiscox, ANSSI (France), etc. Mais cette information reste incomplète, car les entreprises préfèrent cacher les incidents cyber pour ne pas subir une atteinte à l'image.

Dans le cadre de notre étude, nous avons analysé la sensibilité du tarif aux jeux de fréquence suivants **:**

		Global	MED	ADM	RETAIL	BSF
	Fréquence 1	*18%*	*20%*	*15%*	*15%*	*10%*
+10pts	*Fréquence 2*	*28%*	*30%*	*25%*	*25%*	*20%*
-5pts	*Fréquence 3*	*13%*	*15%*	*10%*	*10%*	*5%*
+15pts	*Fréquence 4*	*33%*	*35%*	*30%*	*30%*	*25%*

Figure 64 : Les jeux de fréquences testées

Pour illustrer notre analyse nous présentons le tableau récapitulatif ci-dessous :

	Global	MED	ADM	RETAIL	BSF
Tarif Fr. 1	484 230	386 501	791 439	515 523	364 868
Tarif Fr. 2	748 270	579 752	1 319 064	859 205	729 736
Tarif Fr. 3	352 209	289 876	527 626	343 682	182 434
Tarif Fr. 4	880 290	676 377	1 582 877	1 031 046	912 170
Fr. 1 vs Fr. 2	55%	50%	67%	67%	100%
Fr. 1 vs Fr. 3	-27%	-25%	-33%	-33%	-50%
Fr. 1 vs Fr. 4	82%	75%	100%	100%	150%

Figure 65 : Tableau récapitulatif : sensibilités à la fréquence

Dans le cas d'une diminution de la fréquence de 5 points, le tarif global (tous secteurs confondus) proposé baisserait de 27 %, alors que la hausse de 15 % de la fréquence entraînerait une hausse du tarif global de 82 %.

Vu autrement, nous observons comme attendu un effet de proportionnalité sur le tarif : une évolution de x% de la fréquence résulterait en une évolution de x% du tarif.

Cette analyse met en évidence l'extrême sensibilité d'un tarif d'assurance à la fréquence annuelle des sinistres et justifie l'importance d'étudier cet aspect-là de la sinistralité du risque cyber.

Impact d'une franchise

Il nous a semblé pertinent, au regard des montants de sinistres cyber observés, d'étudier la mise en place de franchises dont nous présentons ici 3 niveaux :

- Franchise 100K€
- Franchise 300K€
- Franchise 500K€

Cette franchise a deux objectifs :

- Lutter contre l'aléa moral et par conséquent responsabiliser les clients ;
- Permettre de diminuer le tarif en partageant le risque.

Une mise en place des franchises permet à l'assureur de proposer un tarif plus compétitif., puisque l'assuré partagera les frais avec l'assureur en cas de sinistre.

Jeu de fréquence 1

			Global	MED	ADM	RETAIL	BSF
		Fréquence	***18%***	***20%***	***15%***	***15%***	***10%***
		Tarif	**484 230**	**386 501**	**791 439**	**515 523**	**364 868**
Sensibilité	Franchise	Franchise 100K	-4%	-5%	-2%	-3%	-3%
		Franchise 300K	-11%	-15%	-6%	-8%	-7%
		Franchise 500K	-18%	-25%	-9%	-13%	-12%
	Plafond	Plafond 2M	-50%	-37%	-67%	-57%	-59%
		Plafond 5M	-21%	-13%	-36%	-22%	-24%
		Plafond 8M	-10%	-5%	-18%	-7%	-11%
	Franchise + Plafond	F 100K - P 2M	-53%	-42%	-69%	-60%	-61%
		F 300K - P 5M	-32%	-28%	-41%	-30%	-31%
		F 500K - P 8M	-27%	-30%	-27%	-20%	-22%

Figure 66 : Impact de la mise en place de la franchise et/ou du plafond

C'est le secteur médical où le tarif diminuerait le plus : le passage de franchise de 100K à 500K se traduirait par une baisse entre 5 % et 25 %, alors que tarif global ne baisserait au-delà de 18%.

A l'inverse, la nature des sinistres du secteur administratif tels que simulés dans notre portefeuille, avec un tarif moyen de base de l'ordre de 800K, donnerait un rôle moindre à la mise en place d'une franchise, dont l'impact tarifaire n'excéderait pas les 9%.

Impact d'un plafond d'indemnisation

Le plafond de couverture est un montant maximal que l'assureur s'engage à payer en cas de sinistre.

L'augmentation du plafond entraîne souvent une augmentation du tarif, car l'assureur supporte un risque plus important.

De la même manière, une diminution du plafond entraînerait une diminution du tarif, mais cette relation n'est pas linéaire.

Il nous a semblé pertinent, au regard des montants de sinistres cyber observés, d'étudier la mise en place de plafonds dont nous présentons ici 3 niveaux :

- Plafond 2M€
- Plafond 5M€
- Plafond 8M€

Pour rappel, nous présentons les résultats de nos analyses dans le tableau ci-dessous :

Jeu de fréquence 1

			Global	MED	ADM	RETAIL	BSF
		Fréquence	***18%***	***20%***	***15%***	***15%***	***10%***
		Tarif	**484 230**	**386 501**	**791 439**	**515 523**	**364 868**
Sensibilité	Franchise	Franchise 100K	-4%	-5%	-2%	-3%	-3%
		Franchise 300K	-11%	-15%	-6%	-8%	-7%
		Franchise 500K	-18%	-25%	-9%	-13%	-12%
	Plafond	Plafond 2M	-50%	-37%	-67%	-57%	-59%
		Plafond 5M	-21%	-13%	-36%	-22%	-24%
		Plafond 8M	-10%	-5%	-18%	-7%	-11%
	Franchise + Plafond	F 100K - P 2M	-53%	-42%	-69%	-60%	-61%
		F 300K - P 5M	-32%	-28%	-41%	-30%	-31%
		F 500K - P 8M	-27%	-30%	-27%	-20%	-22%

Nous observons que la fixation du plafond à 2 millions permet de diminuer le tarif de base de 50% pour le tarif global.

Cette baisse pourrait atteindre 67% pour le secteur administratif, qui, comme évoqué dans la partie précédente, semble structurellement faire face à une sinistralité plus importante, dès lors, la mise en place d'un plafond ramené à 2 millions s'avère particulièrement bénéfique sur le tarif du secteur administratif.

De manière générale, compte tenu de la sévérité des sinistres simulés à partir de la base PRC, la mise en place des plafonds retenus ici semble plus intéressante que les franchises en termes d'économies sur les tarifs d'assurance.

Bien entendu, la contrepartie est que les entreprises victimes d'incidents supérieurs au montant du plafond sont susceptibles de devoir engager des coûts très importants pour leur prise en charge.

Impact d'une combinaison franchise - plafond

Il nous également semblé intéressant d'étudier une formule combinant une franchise ainsi qu'un plafond.

Pour ce faire, nous avons étudiée trois formules :

- Franchise 100K€ - Plafond 2M€
- Franchise 300K€ - Plafond 5M€
- Franchise 500K€ - Plafond 8M€

Pour rappel, nous présentons les résultats de nos analyses dans le tableau ci-dessous :

Jeu de fréquence 1

			Global	MED	ADM	RETAIL	BSF
		Fréquence	***18%***	***20%***	***15%***	***15%***	***10%***
		Tarif	**484 230**	**386 501**	**791 439**	**515 523**	**364 868**
Sensibilité	Franchise	Franchise 100K	-4%	-5%	-2%	-3%	-3%
		Franchise 300K	-11%	-15%	-6%	-8%	-7%
		Franchise 500K	-18%	-25%	-9%	-13%	-12%
	Plafond	Plafond 2M	-50%	-37%	-67%	-57%	-59%
		Plafond 5M	-21%	-13%	-36%	-22%	-24%
		Plafond 8M	-10%	-5%	-18%	-7%	-11%
	Franchise + Plafond	F 100K - P 2M	-53%	-42%	-69%	-60%	-61%
		F 300K - P 5M	-32%	-28%	-41%	-30%	-31%
		F 500K - P 8M	-27%	-30%	-27%	-20%	-22%

Comme attendu, la formule combinée franchise – plafond offre les tarifs les plus compétitifs, lesquels seront bénéfiques aux entreprises non touchées par des incidents cyber.

A l'inverse, les entreprises concernées par de tels tarifs devront soutenir le poids de l'assurance mais également ceux de la franchise et du plafond.

Il peut alors être intéressant de se pencher sur une analyse des « gagnants » et des « perdants » en portefeuille, que nous développerons un peu plus loin dans ce mémoire.

Tarif segmenté vs tarif global

Il est intéressant de déterminer l'impact du secteur d'activité sur le tarif proposé.

En effet, selon plusieurs études, le secteur médical, le secteur financier, le secteur public et le secteur commercial sont dans le top 10 des secteurs les plus touchés par les incidents cyber. Parmi eux, le secteur médical et le secteur financier ont le coût d'une violation de données le plus élevé selon l'étude de l'Institut de Ponemone : 10,1 millions $ le coût moyen d'une violation de donnée dans le secteur médical (USA) et 5,97 millions $ pour le secteur financier.

A défaut de pouvoir se baser sur les mêmes données, nous tirons des observations de l'effet de la segmentation du tarif par secteur d'activité vs. tarif global sur la base des données issues de la base PRC et plus précisément de notre portefeuille simulé.

Nous considérons donc possible que les conclusions tirées ci-après soient biaisées par rapport à celles de l'Institut Ponemone.

IMPACT SEGMENTATION TARIF SECTORIEL							
			Global	MED	ADM	RETAIL	BSF
		Fréquence	***18%***	***20%***	***15%***	***15%***	***10%***
		Tarif	484 230	-20%	63%	6%	-25%
Sensibilité	Franchise	Franchise 100K	466 121	-21%	67%	8%	-24%
		Franchise 300K	431 007	-24%	73%	10%	-22%
		Franchise 500K	398 117	-27%	80%	13%	-19%
	Plafond	Plafond 2M	244 218	0%	6%	-10%	-38%
		Plafond 5M	380 211	-12%	34%	6%	-27%
		Plafond 8M	438 165	-16%	48%	9%	-26%
	Franchise + Plafond	F 100K - P 2M	226 110	-1%	8%	-9%	-38%
		F 300K - P 5M	326 988	-15%	42%	10%	-23%
		F 500K - P 8M	352 052	-23%	63%	18%	-20%

Figure 67 : Impact de la segmentation

Dans le cas d'un tarif sans segmenter, ce sont surtout les entreprises du secteur administratif qui seraient gagnantes.

En effet, leur tarif de base connaitrait une baisse de 63% par rapport à une démarche où les tarifs seraient estimés par secteur. Ce sont donc celles qui sembleraient le plus tirer bénéfice de l'effet de mutualisation au risque inhérent aux autres secteurs d'activité.

Cet effet permettrait également une meilleure mutualisation du risque et potentiellement la mise en place de tarifs plus compétitifs sur l'ensemble du portefeuille.

A l'inverse, les entreprises du secteur médical (hausse de 20%) et du secteur financier (hausse de 25%) seraient celles qui auraient le plus à perdre d'une mutualisation du tarif auprès de l'ensemble des secteurs et qui gagneraient de la segmentation du tarif

Ces dernières seraient sans doute largement favorables à l'application d'un tarif sectoriel.

Nous observons également que la mise en place d'un plafond à 2 millions sur notre portefeuille permettrait d'homogénéiser les tarifs segmentés, à l'exception des entreprises du secteur financier. La mise en place de franchise de 500 000 accompagnée d'un plafond à 8 millions donne un résultat proche du résultat du tarif central.

Analyse des gains et pertes par entreprise en portefeuille

Nous nous sommes également intéresses aussi à la proportion des entreprises qui seraient perdantes ou gagnantes du fait de la politique tarifaire mise en place par l'assureur.

Dans cette partie, nous ne nous intéressons qu'aux entreprises sinistrées en 2023, les autres étant par nature « perdantes » du fait d'avoir payé une prime d'assurance sans la contrepartie d'une mise en jeu des garanties.

	Scénario	Année 2023
GAIN (+) / PERTE (-) ENTREPRISE	Tarif central	10%
	Franchise 100K	14%
	Franchise 300K	22%
	Franchise 500K	29%
	Plafond 2M	5%
	Plafond 5M	8%
	Plafond 8M	9%
	F100 et P2	8%
	F300 et P5	19%
	F500 et P8	27%

Figure 68 : Gain / perte des entreprises

Nous observons que 10% des entreprises seraient perdantes en 2023 avec un tarif « central », c'est-à-dire sans franchise ni plafond.

Par ailleurs, comme nous pouvions l'anticiper, la mise en place d'une franchise impliquant la prise en charge de la première tranche d'indemnisation ne s'avère pas forcément très

avantageuse pour les entreprises (entre 14% et 29% des entreprises sinistrées seraient perdantes). La compétitivité du tarif franchisé ne semble pas permettre de s'y retrouver.

A l'inverse, la mise en place d'un plafond semble donner des résultats plus intéressants pour l'économie des entreprises.

A titre indicatif, le scénario avec la franchise à 2 millions impliquerait que seules 5% des entreprises assurées et sinistrées seraient désavantagées par leur assurance.

Conclusion

Le risque cyber représente un enjeu majeur pour les entreprises aujourd'hui, et la nature évolutive et systémique qui le caractérise rend son appréhension particulièrement complexe.

Par ailleurs, la montée en puissance de l'internet des objets et l'instabilité géopolitique du moment ne font qu'exacerber le besoin de se prémunir contre les impacts potentiels des attaques cyber.

Il est important pour les entreprises, quelle que soit leur taille, de mettre en place des méthodes de prévention adaptées, passant notamment par une formation efficace de leurs salariés, étant donné qu'une part importante du risque cyber incombe aujourd'hui à une erreur liée à l'homme.

Une bonne maîtrise du réseau informatique mis en place par les entreprises paraît aussi être un moyen de se prémunir autant que possible contre les incidents cyber.

Toutefois, le risque zéro n'existe pas et cela se vérifie particulièrement avec le risque cyber, dont les incidents sont difficiles à prévoir, évoluent dans leurs procédés et peuvent avoir de très lourdes conséquences pour des entreprises de plus en plus vulnérables.

Dès lors, il semble particulièrement important pour ces entreprises de se couvrir contre les impacts de potentiels incidents cyber (exemple : violation des données, pertes d'exploitation, frais d'expertises, frais d'avocats, atteinte à la réputation de l'entreprise).

L'enjeu autour de l'assurance cyber est donc particulièrement important, au regard notamment du niveau de couverture actuel d'entreprises, qui semblent en moyenne faiblement prémunies contre les impacts d'éventuels incidents.

Si les grandes entreprises présentent un taux de couverture intéressant (87%), ces dernières paraissent malgré tout insuffisamment couvertes, en atteste le S/P observé en France en 2020 qui témoigne d'une sinistralité difficile à estimer (S/P = 190%)

Quant aux PME et ETI, leurs taux de couverture actuels (2020) sont préoccupants (respectivement 0,0026% et 8%).

L'une des raisons derrière ces observations réside dans le fait que l'estimation du risque cyber s'appuie sur un faible volume de données, rendant le risque cyber particulièrement difficile à évaluer.

La base PRC utilisée dans le cadre de ce mémoire présente des limites significatives pour la modélisation du risque cyber :

- *Portée géographique limitée* : la base PRC recense essentiellement des violations de données qui se sont produites aux Etats-Unis. L'état de la menace cyber aux Etats-Unis peut ne pas refléter les spécificités de la menace cyber en France : contexte géopolitique, les vulnérabilités spécifiques, les pratiques en matière de cybersécurité, …
- *Différences légales et réglementaires* entre la France et les Etats-Unis : Exemple : la prédominance du secteur médical depuis 2010 dans la base PRC pourrait s'expliquer par la modification du processus de notification des sinistres cyber dans le domaine médical.
- *Sous-déclaration des sinistres* : la base PRC ne reflète pas la situation réelle concernant le nombre de fuites observées, il y a beaucoup plus d'incidents réellement survenus que

d'incidents déclarés. Les entreprises sont réticentes à déclarer des fuites de données pour éviter une mauvaise publicité.

- *Incertitude sur les impacts financiers* des incidents cyber : il est difficile d'évaluer précisément le coût des sinistres. La base PRC possède une variable Total.Records qui indique le nombre d'enregistrements compromis par sinistre, mais sans donner l'information quant à son coût réel.
- *Nature évolutive de la sinistralité cyber* : Les modèles doivent être constamment mis à jour pour suivre l'évolution du risque. Par conséquent, les résultats pourraient devenir rapidement obsolètes si la base PRC n'est pas mise à jour de manière régulière.
- *Données incomplètes* : présence des valeurs manquantes, des variables pertinentes absentes de la base PRC (taille de l'entreprise, …)

Il y a là un réel enjeu dans la construction d'une base de données recensant les incidents cyber au niveau européen et/ou français, qui serait plus pertinente et permettrait de recenser des variables non présentes dans la base PRC, comme le coût par sinistre, la taille de l'entreprise ou encore son chiffre d'affaires, permettant ainsi une meilleure prédiction du risque.

La base PRC reste la base publique la plus complète à ce jour. Par conséquent, en nous appuyant sur les données de la table PRC, qui recense des informations sur les incidents cyber survenus aux Etats Unis depuis plusieurs décennies, nous avons essayé, dans ce mémoire, de mettre en évidence des méthodes permettant une meilleure compréhension des données sous-jacentes et de permettre la mise en place de modèles de prédiction efficaces du nombre de sinistres ainsi que du nombre d'enregistrements compromis par incident (Partie 2).

Pour ce faire, une analyse comparative des méthodes de Séries Temporelles et de Machine Learning a été réalisée.

Nous pouvons observer que les méthodes de Machine Learning se sont avérées plus performantes pour la prédiction de la sinistralité cyber que les méthodes de Séries Temporelles.

Les modèles de régression testés initialement ne donnant pas de résultats probants ($R2<0.05$), nous avons donc procédé à un retraitement de la base pour pouvoir utiliser, dans les modèles de régression, les données mensuelles en fonction du secteur et de la zone géographique. Ce procédé a permis d'améliorer sensiblement les performances des modèles. Même les modèles les moins performants donnaient des résultats intéressants.

Parmi ces modèles, c'est le modèle de réseaux de neurones s'est révélé particulièrement performant dans la prédiction du nombre d'incidents mensuels (à savoir pour l'estimation de la fréquence : nombre de sinistres). Concernant la sévérité des incidents, tous les modèles ont montré des performances relativement proches.

Nous considérons que la performance des modèles de neurones peut être significativement améliorée en déterminant une meilleure calibration. Cependant en gagnant en précision, on « gagne » également en complexité, car il s'agît d'une boite noire.

Les autres modèles testés présentent l'avantage d'être plus simples et plus rapides à calibrer pour des performances sensiblement équivalentes.

Une piste d'amélioration consisterait à incorporer l'information sur la taille de l'entreprise et/ou son chiffre d'affaires. Cela permettrait de construire des modèles plus performants.

Dans nos travaux, nous rappelons que la variable cible a subi une transformation logarithmique dans l'objectif d'améliorer les performances des modèles.

Il s'agît d'un point particulièrement sensible, car si les modèles mis en place produisent des résultats relativement satisfaisants sur l'estimation du logarithme de la gravité, la transformation inverse (exponentielle) se traduit par des écarts importants, comme en atteste la prédiction faite sur 2023 dans le cadre de l'analyse tarifaire (Partie 4).

Également, nous avons porté notre attention sur l'étude des dépendances entre les attaques cyber, toujours en se basant sur les données issues de la table PRC (Partie 3).

L'objectif était d'identifier la structure de dépendance entre :

- les différentes zones géographiques,
- les différents types de violations de données
- les différents types de secteurs d'activité

En dernier lieu, nous avons proposé une analyse tarifaire pour une garantie contre la violation de données à destination d'entreprises (Partie 4).

En se basant sur un portefeuille fictif issu de la base PRC, nous avons notamment observé l'intérêt d'un tarif global en comparaison d'un tarif sectoriel. Celui-ci permet, par une plus grande mutualisation avec de « bons risques », de rendre plus accessible la couverture des entreprises issues de secteurs les plus sensibles.

Par ailleurs, après avoir testé différentes formules d'atténuation des montants assurés, il en ressort que la mise en place d'un plafond à 2 millions d'€ donne, sur notre portefeuille simulé, des résultats intéressant avec un tarif global presque divisé par deux.

L'intérêt de cette formule semble corroboré par l'analyse des gains et pertes des entreprises sinistrées.

En effet, le plafond à 2 millions d'€ fait ressortir que seules 5% des entreprises assurées et sinistrées seraient perdantes par l'effet de cette formule d'assurance.

Par la suite, il pourrait être intéressant de mener une analyse plus poussée afin d'identifier le niveau de plafond qui optimiserait la couverture.

Par ailleurs, le niveau globalement élevé des polices d'assurance estimées ici tend à témoigner du faible niveau de TPE et PME assurées contre le risque cyber aujourd'hui.

Si le niveau des tarifs présentés dans le cadre de notre analyse semble acceptable pour de grands groupes, il l'est moins pour les plus petites structures de notre économie.

Dès lors, la prise en compte de variables supplémentaires comme la taille des entreprises ou leur chiffre d'affaires, pourrait représenter une piste intéressante pour une mutualisation plus juste du risque au sein du portefeuille.

Table des figures

Bibliographie

Cyber attaques en 2022 : tous les signaux sont au rouge pour toutes les entreprises - La Revue du Digital

rapport_45_f.pdf (banque-france.fr)

https://www.globalsecuritymag.fr/Barometre-des-risques-2022-d,20220118,120855.html

https://www.cesin.fr/actu-7eme-edition-du-barometre-annuel-du-cesin-enquete-exclusive-sur-la-cybersecurite-des-entreprises-francaises.html

https://www.clubic.com/antivirus-securite-informatique/actualite-409014-cybercriminalite-la-russie-a-genere-les-trois-quarts-des-revenus-lies-aux-attaques-par-ransomware.html

https://blogs.microsoft.com/on-the-issues/2022/04/27/hybrid-war-ukraine-russia-cyberattacks/

Montée en puissance des cyberattaques : il est temps pour les TPE et PME de prendre conscience du risque (journaldunet.com)

L'enquête LUCY (LUmière sur la CYberassurance) de l'AMRAE (2021)

L'enquête LUCY (LUmière sur la CYberassurance) de l'AMRAE (2022)

https://www.cyber-cover.fr/cyber-documentation/rgpd/assurance-cyber-rgpd-pourquoi-sassurer

https://www.cnil.fr/sites/default/files/atoms/files/cybersecurite_-_chiffres_2021.pdf

https://www.cnil.fr/fr/definition/violation-de-donnees

https://www.apref.org/wp-content/uploads/2016/07/note_apref_cyber_risque-1.pdf

Le risque cyber dans le secteur financier | Direction générale du Trésor (economie.gouv.fr)

https://www.c-risk.com/fr/blog/gestion-des-risques/

https://www.avg.com/fr/signal/what-is-malware

https://www2.deloitte.com/fr/fr/pages/risque-compliance-et-controle-interne/articles/cyberattaques-chiffrer-les-impacts.html

https://www.cgi.com/sites/default/files/2019-07/cgi-understanding-cybersecurity-standards-white-paper-fr.pdf

https://www.historyhit.com/the-biggest-cyberattacks-in-history/

https://www.usine-digitale.fr/article/il-est-urgent-de-trouver-une-solution-aux-problemes-d-assurance-cyber.N1779877

https://www.cri4data.com/index.php/2018/11/16/comment-les-assureurs-tarifieraient-ils-les-produits-cyber-avec-peu-de-donnees/ (pas utilise à voir pour les copules)

https://www.ibm.com/fr-fr/topics/decision-trees

Poterie, Audrey. (2018). Arbres de décision et forêts aléatoires pour variables groupées. (figure arbre de décision)

https://blog.ysance.com/algorithme-n2-comprendre-comment-fonctionne-un-random-forest-en-5-min

https://aws.amazon.com/fr/what-is/boosting/

https://mrmint.fr/naive-bayes-classifier

https://www.geeksforgeeks.org/naive-bayes-classifiers/

Icon Schematic Processing Data Inside Neural Stock Vector (Royalty Free) 1130525327 | Shutterstock

www.shutterstock.com (image neural network)

https://meritis.fr/deep-learning/

https://medium.com/france-school-of-ai/math%C3%A9matiques-des-r%C3%A9seaux-de-neurones-code-python-613d8e83541

https://www.headmind.com/fr/text-mining-classification-automatique-de-textes/

Antoine Sainson, Hugo Linsenmaier, Alexandre Majed, Xavier Cadet, Abdessalam BouchekiLSE : DEFT 2018 : Classification de tweets basée sur les réseaux de neurones profonds

François Husson, R pour la statistique

https://datascientest.com/series-temporelles#:~:text=Math%C3%A9matiquement%20une%20s%C3%A9rie%20temporelle%20c,c'est%20pr%C3%A9dire%20le%20futur.

Gildas Mazo : Construction et estimation de copules en grande dimension

Belguise Olvie, Charles Levi : Tempêtes : Etude des dépendances entre les branches Automobile et Incendie à l'aide de la théorie des copulas Topic 1 Risk evaluation

Bedi Nathalie : Modélisation du risque de tempête en France métropolitaine

Charpentier : Copules

Martin Eling , Kwangmin Jung : Copula approaches for modeling cross-sectional dependence of data breach losses

Sébastian Farkas, Olivier Lopez, Maud Thomas :Cyber claim analysis using Generalized Pareto regression trees with applications to insurance

Yannick Bessy-Roland : Modélisation stochastique individuelle de sinistres cyber

https://gestiondesrisques.net/2022/03/30/ca-bouge-du-cote-du-cyber-risque

https://www.c-risk.com/fr/blog/gestion-des-risques

https://www.tresor.economie.gouv.fr/Articles/2021/12/14/le-risque-cyber-dans-le-secteur-financier

http://pre.fmglobal-touchpoints.fr/se-demarquer/cyber-risques-une-realite-a-prendre-en-compte

https://www.gouvernement.fr/risques/cybercriminalite

https://www.la-vie-nouvelle.fr/infos/dossiers/se-premunir-des-cyberattaques

https://www.cgi.com/sites/default/files/2019-07/cgi-understanding-cybersecurity-standards-white-paper-fr.pdf

https://www.larousse.fr/encyclopedie/divers/ISO/125281

https://www.cyber-cover.fr/cyber-documentation/rgpd/assurance-cyber-rgpd-pourquoi-sassurer

https://www.cnil.fr/sites/default/files/atoms/files/cybersecurite_-_chiffres_2021.pdf

https://www.cgi.com/sites/default/files/2019-07/cgi-understanding-cybersecurity-standards-white-paper-fr.pdf

https://www.touteleurope.eu/economie-et-social/cybersecurite-que-fait-l-union-europeenne

https://trustpair.fr/blog/cybersecurity-act-la-loi-europeenne-sur-la-cybersecurite

https://www.touteleurope.eu/economie-et-social/cybersecurite-que-fait-l-union-europeenne

https://www.ibm.com/fr-fr/topics/decision-trees

https://projeduc.github.io/intro_apprentissage_automatique/arbres.html

https://mrmint.fr/naive-bayes-classifier

https://actualiteinformatique.fr/cybersecurite/definition-malware

https://www.ibm.com/fr-fr/topics/decision-trees

Annexes

State	Total records
Alabama	77
Alaska	26
Arizona	111
Arkansas	83
California	1325
Colorado	173
Connecticut	149
Delaware	21
Florida	451
Georgia	251
Hawaii	28
Idaho	24
Illinois	336
Indiana	213
Iowa	66
Kansas	54
Kentucky	118
Louisiana	61
Maine	32
Maryland	342
Massachusetts	248
Michigan	152
Minnesota	145
Mississippi	36
Missouri	143
Montana	35
Nebraska	43
Nevada	65
New Hampshire	44
New Jersey	160
New Mexico	54
New York	615
North Carolina	210
North Dakota	10
Ohio	262
Oklahoma	68
Oregon	128
Pennsylvania	274
Rhode Island	41
South Carolina	72
South Dakota	13
Tennessee	162
Texas	577
Utah	61
Vermont	32
Virginia	199
Washington	195
West Virginia	29
Wisconsin	101
Wyoming	16
Total général	**8131**

Type of breach	Nombre
CARD	67
DISC	1803
HACK	2427
INSD	589
PHYS	1692
PORT	1133
STAT	245
UNKN	175
Total	8131

Type d'organisation	NB
BSF	732
BSO	996
BSR	606
EDU	838
GOV	675
MED	4175
NGO	109
Total général	**8131**

Année	Nombre d'incidents
2005	133
2006	456
2007	442
2008	348
2009	260
2010	759
2011	748
2012	848
2013	819
2014	834
2015	532
2016	794
2017	669
2018	418
2019	71
Nombre d'incidents	**8131**