Analyse d'événements Cyber grâce à un arbre de régression construit selon une vraissemblance de pareto généralisée et application à la tarification pour l'assurance des risques Cyber

Sébastien Farkas [1] & Olivier Lopez [2] & Maud Thomas [3]

[1] *Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, sebastien.farkas@sorbonne-universite.fr*

[2] *Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, olivier.lopez@sorbonne-universite.fr*

[3] *Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, maud.thomas@sorbonne-universite.fr*

Résumé.

Nous proposons une méthodologie d'analyse des bases de données de sinistres Cyber qui considère l'hétérogénéité des données pour proposer une calibration des tarifs et des conséquences d'un évènement extrême liés à un portefeuille de polices d'assurances Cyber. Nous appliquons cette méthodologie à une base publique constituée par l'association "Privacy Rights Clearinghouse" qui répertorie des événements de failles de données concernant des citoyens américains. Nous apportons une attention particulière aux valeurs extrêmes et analysons l'hétérogénité des données en réalisant une classification hiérarchique inspirée des arbres de régression par maximum de vraissemblance. Nous étayons cette modélisation de la sévérité par une modélisation de la fréquence des sinistres pour construire un modèle de tarification simple applicable à l'assurance du risque Cyber.

Mots-clés. Modèles pour l'assurance et la finance, Valeurs extrêmes, ...

Abstract.

In this paper we propose a methodology to analyze the heterogeneity of cyber claims databases which can then be applied to calibrate the pricing and the extreme scenarios of cyber risk. We focus on data breaches events considering the Privacy Rights Clearinghouse database, a public database which is considered as a benchmark for data breaches analysis. Using regression trees, we investigate the heterogeneity of the reported data breaches. A particular attention is devoted to the tail of the distribution, adapting the construction of Maximum Likelihood Regression Trees (MLRT) with a Generalized Pareto likelihood as splitting criterion. Combining this analysis with a model for the frequency of the claims, we develop a simple pricing model for cyber insurance.

Keywords. Models for insurance and finance, Extreme Value Theory, ...

1 Les données et les risques cyber

1.1 Introduction aux risques cyber

Les risques Cyber peuvent être définis comme l'ensemble des risques émanant de l'utilisation de données électroniques et de leur transmission. Beaucoup d'entreprises ont saisi l'informatique comme une opportunité de développement et font ainsi face aux riques Cyber. Le marché de l'assurance propose depuis peu des produits permettant aux sociétés de s'assurer contre l'occurence d'événements Cyber. La détermination des prix et la gestion du risque liés à ces nouveaux produits ne pouvant se faire par une étude de l'historique des sinistres, les assureurs calibrent principalement leurs prix grâce à des avis d'experts et des analyses de données macroéconomiques.

1.2 L'hétérogénéité des données

Toutefois, la commercialisation de ces produits permet la constitution de bases de données sinistres. Ces données présentent des caractéristiques propres aux risques Cyber et soulèvent autant de problématiques. Nous considérons la problématique lié à l'hétérogénéité des données, conséquence de l'évolution des risques Cyber mais aussi de l'évolution de la qualité et des sources de données.

2 Méthodologie proposée

2.1 Modélisation

Nous proposons une modélisation de la sévérité des sinistres cyber qui se concentre sur l'étude des valeurs extrêmes et qui prend en compte l'hétérogénéité des données. Nous construisons un arbre de régression inspiré des arbres de régression par maximum de vraissemblance introduits par [3]. En effet, le critère de coupe de l'arbre de régression proposé dépend de la vraissemblance des données à une loi de pareto généralisée. Nous construisons ainsi des groupes homogènes d'événements extrêmes qui nous permettent de modéliser la sévérité. Pour compléter cette étude, nous réalisons une étude de la fréquence des sinistres. En combinant ces deux approches, nous aboutissons à une proposition simple d'un modèle de tarification.

2.2 Application

Nous appliquons notre méthodologie à une base de données publique qui répertorie des événements de failles de données concernant des citoyens américains: la base de l'association "Privacy Rights Clearinghouse". Cette base est une référence pour la calibration de modèles mathématiques pour l'assurance des risques cyber. En effet, [2] et

[4] l'ont analysé pour étudier le phénomène des failles de données et [1] a proposé une analyse actuarielle de cette base aboutissant à une tarification.

Bibliographie

[1] Edwards B., Hofmeyr S. and Forrest S. (2016), "Hype and heavy tails: A closer look at data breaches." *Journal of Cybersecurity*, vol. **2** (2057-2085), pp. 3-14.

[2] Eling M. and Loperfido N. (2017), "Data breaches: Goodness of fit, pricing, and risk measurement." *Insurance: Mathematics and Economics*, vol. **75** (0167-6687), pp. 126-136.

[3] Su X., Wang M. and Fan J. (2004), "Maximum Likelihood Regression Trees." *Journal of Computational and Graphical Statistics*, vol. **13** (1061-8600), pp. 586-598.

[4] Wheatley S., Maillart T. and Sornette D. (2016), "The extreme risk of personal data breaches and the erosion of privacy." *European Physical Journal B*, vol. **89** (1434-6036), pp. 7.