Extreme value regression and insurability of natural disasters

Advisor Olivier Lopez, CREST ENSAE IP Paris
Funding Chair of Excellence CARE (Allianz - Ensae - Fondation du Risque)

Keywords: Machine learning, extreme value theory, geographic data, insurance.

1 Motivation

The resurgence of natural disasters is a corollary of climate change. Insurance mechanisms are an important component of adapting to this new context because the financial protection they offer must absorb a significant portion of the economic burden associated with reconstruction. Additionally, insurance also serves as a vector for the dissemination of prevention standards, which has historically been one of its roles (for example, the development of fire standards was largely promoted by British insurance companies after the Great Fire of London in 1666).

However, faced with the increase in both the frequency and severity of these natural disasters, the question of insurability of these risks arises. At the heart of this concept of insurability lies the principle of mutualization. This principle is essentially a translation of the law of large numbers and the central limit theorem in the insurance context: if the losses associated with each insured in a portfolio are independent identically distributed random variables, the loss recorded by the insurer becomes almost certain (approximately the expected value of a loss multiplied by the number of insured, known as the central scenario), with some uncertainty depending on the variance of the phenomenon, but which is negligible compared to the amount represented by the central scenario. This vision is asymptotic, thus requiring a sufficient volume of insured individuals.

The aim of this PhD thesis is then twofold:

1. **fit statistical models to analyze the severity of catastrophic events on an insurance portfolio, based on meteorological, satellite, socio-economical data, relying on machine learning and extreme value theory.**

2. **use this model to project the evolution of a portfolio and study the viability of risk sharing strategies (reinsurance, public-private strategies...) and prioritize prevention.**

The project is part of a Chair of Excellence launched by Allianz, Ensae, Fondation du Risque in December 2024[1]. The PhD candidate will work in close interaction with the teams of these three institutions involved in the project, and will be directly connected with business cases and advanced research topics.

2 Extreme value theory

In the case of natural disasters, the random variable representing the distribution of losses has a heavy tail due to the extreme volatility of damage costs. It is often modeled by a variable Y from Fréchet's domain, that is defined by a survival function of the type

$$P(Y \geq t) = \frac{l(t)}{t^{1/\gamma}}, \tag{1}$$

where l is a slow-varying function, and $\gamma > 0$.

This parameter γ, also known as the tail index, plays a particular role as it describes the shape of the distribution tail, involved in large-scale losses. Pickands–Balkema–De Haan Theorem (see [Pickands III, 1975]) indeed shows that the distribution tail of a random variable Y of the type (1) can be approximated by a Generalized Pareto distribution whose shape parameter is guided by this tail index. Furthermore, it can be seen that if γ is greater than 1, the variable has no expectation. In this case, we encounter the most obvious situation of uninsurability: the law of large numbers does not apply, and mutualization cannot operate. Even if γ is strictly less than 1, the situation is all the more complicated when its value is high, which can again pose an insurability problem: the insurance premium, which more or less reflects the risk expectation, can become too high, dissuading insured parties from taking out a policy (which again leads to a loss of mutualization for the insurer since it deviates from the asymptotic

[1] https://newsroom.allianz.fr/allianz-france-la-fondation-du-risque-et-lensae-paris-lancent-une-chaire-sur-lassurabilite-des-risques-emergents/

regime), or the insurer will establish restrictive conditions (indemnity limits, deductibles...) that restrict the scope of compensation.

If we consider an insured party (or a geographical area) with characteristics $X \in R^d$, the tail index of the conditional distribution of Y given $X = x$ is a function $\gamma(x)$. Knowing this function is important as it will allow for optimizing the terms of risk coverage. Indeed, not understanding the impact of characteristics X influences $\gamma(X)$, leading to considering the insured population as an undifferentiated mixture of heavy-tailed distributions, where the gamma index is the infinity norm of the $\gamma(x)$ function, which, in extreme cases, leads to considering all risks as uninsurable when some of them could be insurable under better conditions.

Another obstacle to mutualization is the breakdown of the assumption of independence between insured parties on which the law of large numbers relies. Indeed, one of the characteristics of natural disaster events is the possibility of simultaneous failure of a large number of insured parties due, in particular, to their geographical proximity. Therefore, attention will also be paid to modeling this dependence.

3 Machine learning for tail regression

Many recent contributions have considered the question of combining machine learning with extreme value analysis, see for example [Velthoen et al., 2023], [Richards and Huser, 2022], [Pasche and Engelke, 2022], [Allouche et al., 2024]. Most of them focus on extreme quantiles, but in the present work, we want to estimate the function $x \to \gamma(x)$ because of the practical interpretation of this parameter. In [Farkas et al., 2021] and [Farkas et al., 2024], regression trees have been used to estimate $\gamma(x)$. These models have been applied to the context of cyber risk analysis, and of natural disasters insurance. However, as it is the case for regression tree approaches, the procedure may not always be stable, in the sense that a small change in the database used to fit the model may lead to a significant change in the fitted model. In this PhD thesis, the idea will be to explore more advanced models, extending the approach of [Farkas et al., 2024] to random forests, gradient boosting or neural networks. Although these procedures ultimately produce "black boxes," it is expected to estimate more precisely the function $x \to \gamma(x)$.

The additional problem that we want to tackle is the presence of dependence between policyholders (or geographic areas that may be stroke by a natural event). Two possible ways of dealing with this dependence could be considered within this PhD thesis, depending on which may seem to be the more appropriate considering the practical cases that are considered. First of all, the copula approach (see [Durante and Salvadori, 2010], [Alidoost et al., 2019]) is a classical way to apprehend multivariate extreme value analysis. Alternatively, the link with graph theory may be done to measure the impact of a network structure (hydrological network, geographic proximity between areas...).

4 Risk transfer for natural disasters coverage

Being able to estimate $\gamma(x)$ is key to optimize risk sharing. Based on this knowledge of the tail of the distribution, it is possible to determine how reinsurance (see [Albrecher et al., 2017]) can be used to absorb claims above some threshold, before considering eventually transfer a part of the risk to the State.

In this perspective, the French "Catastrophes Naturelles" regime (see [Cazaux et al., 2019], [Barry, 2024], [Heranval et al., 2023]) is a special case of risk transfer mechanism based on a public / private partnership. To fluidify the market of insurance against natural disasters, the State offers an unlimited reinsurance guarantee (under conditions) if the insurance capacity is insufficient. The increase in both frequency and severity of natural disasters can jeopardize the viability of such regimes.

The evolution of the risk can indeed create some difficulties for the policyholder to find insurance at a reasonable price, and create distortion on the territory. The analysis of the tail of the distribution is essential to define strategies that are compatible with the principle of solidarity.

The tools developed in this PhD thesis will contribute to discuss evolution of such regimes mixing insurance, reinsurance and public intervention.

5 Skills

The PhD candidate should satisfy the following requirements:

- master degree in applied mathematics, ideally statistics or probability;
- computer skills in Python and/or R, with the ability to work with large geographical databases;
- good level of English;

- affinity with insurance and economic evaluation of disasters, ideally (but not mandatory) with basic knowledge in actuarial science.

References

[Albrecher et al., 2017] Albrecher, H., Beirlant, J., and Teugels, J. L. (2017). *Reinsurance: actuarial and statistical aspects*. John Wiley & Sons.

[Alidoost et al., 2019] Alidoost, F., Su, Z., and Stein, A. (2019). Evaluating the effects of climate extremes on crop yield, production and price using multivariate distributions: A new copula application. *Weather and climate extremes*, 26:100227.

[Allouche et al., 2024] Allouche, M., Girard, S., and Gobet, E. (2024). Estimation of extreme quantiles from heavy-tailed distributions with neural networks. *Statistics and Computing*, 34(1):12.

[Barry, 2024] Barry, L. (2024). The moral economies of natural disasters insurance: solidarity or individual responsibility? *Journal of Cultural Economy*, 17(1):39–54.

[Cazaux et al., 2019] Cazaux, E., Meur-Férec, C., and Peinturier, C. (2019). Le régime d'assurance des catastrophes naturelles à l'épreuve des risques côtiers. aléas versus aménités, le cas particulier des territoires littoraux. *Cybergeo: European Journal of Geography*.

[Durante and Salvadori, 2010] Durante, F. and Salvadori, G. (2010). On the construction of multivariate extreme value models via copulas. *Environmetrics: The official journal of the International Environmetrics Society*, 21(2):143–161.

[Farkas et al., 2024] Farkas, S., Heranval, A., Lopez, O., and Thomas, M. (2024). Generalized pareto regression trees for extreme events analysis. *To appear in Extremes*.

[Farkas et al., 2021] Farkas, S., Lopez, O., and Thomas, M. (2021). Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, 98:92–105.

[Heranval et al., 2023] Heranval, A., Lopez, O., and Thomas, M. (2023). Application of machine learning methods to predict drought cost in france. *European Actuarial Journal*, 13(2):731–753.

[Pasche and Engelke, 2022] Pasche, O. C. and Engelke, S. (2022). Neural networks for extreme quantile regression with an application to forecasting of flood risk. *arXiv preprint arXiv:2208.07590*.

[Pickands III, 1975] Pickands III, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131.

[Richards and Huser, 2022] Richards, J. and Huser, R. (2022). A unifying partially-interpretable framework for neural network-based extreme quantile regression. *arXiv preprint arXiv:2208.07581*.

[Velthoen et al., 2023] Velthoen, J., Dombry, C., Cai, J.-J., and Engelke, S. (2023). Gradient boosting for extreme quantile regression. *Extremes*, 26(4):639–667.