Université de Strasbourg
UFR de Mathématique Informatique

Volatility features in Frequency-Severity Catastrophe models with application of Generalized Linear Models and Multifractal theory

Application of derivative-reinsurance instruments

Kristiano BEJKO

Tuteurs académiques : Alexandre YOU, Jean-Philippe BOUCHER
Tuteur professionnel à la CCR : Laurent IMBERT

Mémoire de master présenté devant le jury de soutenance IA

duas
Diplôme Universitaire des **Actuaires de Strasbourg**

CCR

"You have your way. I have my way. As for the right way, the correct way, and the only way, it does not exist."

Friedrich Nietzsche in Thus spoke Zarathustra

Acknowledgements

In order to conceive, nourish and finish my master's thesis, several were they that did not hesitate to offer their limitless help.

I would like to deeply thank Pr. Dr. Jean-Philippe BOUCHER, for without his help, this thesis wouldn't have a subject to treat. From the beginning in May 2016, when he proposed me a subject based on one of his scientific papers, till now, I have very much appreciated his remarks and commentaries that made me construct and improve the models.

I vividly address my acknowledgements to Alexandre YOU, my academic supervisor. Thanks to his rich remarks on every section, I have no doubts about the exactness and the persnickety of the theoretical models and their applications.

I am also very thankful to Alpha BAH, manager of the Non-Life Insurance pole of the Open Market Reinsurance department, Laurent IMBERT, my professional supervisor and certified actuary, and to the team of the same pole, who allowed me to do the last year internship at Caisse Centrale de Réassurance - CCR. My manager and professional supervisor were always available when I needed their point of view on reinsurance applications in my thesis.

I wouldn't miss to mention Georges GUZMAN, Cat modeler at CCR, who showed me a wider vision of Cat models currently used and helped me figure out the limits and perspectives of my model.

Last but not least, I would like to show my gratitude towards Mario ZYLA, a close friend of mine, whose help regarding the English vocabulary and grammar was essential.

Abstract

The cost of natural catastrophes plays a major role in the (re)insurance market. Given the impact of climatic and socio-economic factors, we observe a continuous increase of this cost. On one hand, global climate change explains partially this trend. On the other hand, population growth and a greater wealth accumulation in urban areas generate higher costs. In order to have a better insight of this topic, we decided to study tornado risk. It is characterized by a high frequency variation and great amounts of costs. It must be noted that the aforementioned trend can be observed in the case of the tornado peril too.

At first, we thought it necessary to present different statistical measures performed on the tornado database, under a time series framework. As it is in most of the cases with time series, the tornado data incorporate volatility at "three levels": seasonality/cyclicality, time trend, and high volatility events, also called jumps. We did take them into account in the probabilistic modeling.

As an actuarial researcher, we had several questions in mind:

- Which models are proposed by university presses?
- How can we take into account volatility measurement?
- Can one conceptualize a product that covers well tornado losses?

In the case of most commonly applied models, the frequency component is modeled separately from the claim severity component. Also, an assumption about the independence between arrivals and claims is made. Nonetheless, these Frequency-Severity models in their classical version are not able to tackle the volatility issue. We decided to use the Generalized Linear Models (GLMs) framework, because it allowed us to express expected mean values (arrival count and claim cost averages) in terms of parameters of seasonality and/or time trend. Poisson-Gamma distributions were chosen for the Frequency-Severity model and were modified with the help of GLMs.

The chosen Poisson-Gamma model with seasonality and/or time trend weakly captures the fine behavior of high volatility. It is in a research paper that we encountered a dynamic model based on some features that can incorporate the three volatility levels at the same time. The features are multifractal processes, or simply multifractals. It basically means the following: For a time series value that varies in the time interval $[t, t + \Delta t]$, the change in value depends on the time interval variation Δt. Thanks to their mathematical complexity, multifractals can inculcate climatic and socio-economic factors in their structure. In this thesis, we studied the theory of multifractals, confirmed their statistical efficiency, and applied them to the tornado database.

In order to compare between the Multifractal model and the Frequency-Severity model with seasonality and/or time trend, we realized a numerical application. A derivative-reinsurance treaty was constructed and its estimated price was calculated by using simulated backtesting data from both models. We also showed the advantage of using

multifractals. Even though there currently is no market price or benchmark reference for such a treaty, the prediction error was calculated for both models and confirmed us the quality of prediction for the Multifractal model.

The goal of our thesis is to help (re)insurers better model tornado risk. One ought to bear in mind that there already exist commercialized/Vendor catastrophe models that also model the risk in question. We conclude our thesis with several pros and cons of the Multifractal model and its comparison with Vendor models.

Key words:

Natural catastrophes, trend, volatility, time series, Cat models, multifractals, reinsurance

Résumé

Le coût des catastrophes naturelles chez les assureurs et les réassureurs a toujours été un enjeu majeur. Suite à de nombreux facteurs climatiques et socio-économiques, ce coût n'a cessé de croître au cours du temps. Cette tendance s'explique d'une part par le dérèglement climatique observé et d'autre part par la hausse et l'urbanisation de la population. Pour mieux appréhender ce sujet, nous nous focaliserons sur l'étude des tornades terrestres. Ce phénomène climatique a une fréquence élevée et cause énormément de dommages chaque année aux Etats-Unis. Au cours des trois dernières décennies, nous avons pu observer une nette tendance à la hausse en termes de fréquence et coût des tornades.

Dans un premier temps, nous avons formalisé les différents éléments présents dans la base de données des tornades sous forme de séries temporelles. Comme c'est le cas sur la plupart des séries temporelles, ces données incorporent une volatilité à « trois niveaux ». Nous avons une saisonnalité, une tendance (dite « trend ») et les événements extrêmes (dits les « sauts »). Nous avons par la suite introduit ces éléments dans la modélisation probabiliste.

A ce premier stade de l'étude, nous nous sommes interrogés sur plusieurs points :

- Quels sont les modèles proposés par le milieu académique ?
- Dans quel cadre théorique pouvons-nous tenir compte de la volatilité ?
- Peut-on commercialiser un produit qui couvre les pertes des tornades ?

Les modèles les plus couramment utilisés traitent séparément la fréquence et la sévérité des sinistres. De plus, ils retiennent l'hypothèse d'indépendance entre le nombre de tornades et des sinistres qu'ils engendrent. Or, ces modèles dans leur version initiale s'avèrent insuffisants pour répondre à notre problématique de volatilité. Pour remédier au problème, nous avons décidé d'utiliser le cadre des Modèles Linéaires Généralisés (GLM en anglais) qui nous a permis d'avoir une relation entre la valeur à prédire (nombre moyen ou coût moyen de tornades) et les paramètres de saisonnalité et de tendance. Pour ce faire, le couple des lois probabilistes Poisson-Gamma a été choisi et a été ensuite soumis au cadre des GLM.

Nous étions conscients que les modèles avec saisonnalité et/ou tendance incorporent bien deux niveaux de volatilité parmi trois, les sauts extrêmes n'étant pas complètement atteints. Dans la littérature universitaire, nous avons rencontré une méthodologie théorique : les processus multifractals. Ils permettent d'inclure les trois niveaux de volatilité simultanément. Par le terme multifractal, nous entendons la relation suivante : les séries temporelles dont la variation de la valeur à l'instant t dépend de l'intervalle de temps Δt. Grâce à la complexité mathématique de ces processus, les facteurs climatiques et socio-économiques, moteurs générateurs des tornades, ont aussi pu être pris en compte dans la modélisation. Dans ce mémoire, nous avons étudié la théorie des multifractals, et avons montré leur efficacité statistique dans une application sur notre base de données.

Pour pouvoir comparer le modèle Multifractal et le modèle de Fréquence-Sévérité avec saisonnalité et/ou tendance, nous avons finalement mis en place une application numérique. Un produit dérivé de réassurance non-proportionnelle a été créé et son prix a été estimé à l'aide des données simulées en backtesting par les deux modèles. L'avantage d'utiliser les multifractals a été mis en avant. Bien qu'il n'existe pas de prix marché pour le produit dérivé que nous avons créé, nous avons donné des éléments sur la qualité de prédiction et la solidité du modèle choisi.

Le but de ce mémoire est d'aider les compagnies de réassurance à mieux modéliser le risque de tornades. Toutefois, il existe sur les marchés les modèles Vendor Cat commercialisés qui modélisent déjà ce risque. Le modèle Multifractal présente plusieurs avantages et inconvénients par rapport aux modèles Vendor que nous avons listés à la fin de ce mémoire.

Mots-clés :

Périls naturels, tendance, volatilité, séries temporelles, modèles Cat, multifractals, réassurance

Table of Contents

Acknowledgements **4**

Abstract **5**

Résumé **7**

List of tables **11**

List of figures **12**

Introduction **13**

1. Natural catastrophes in the US **16**

1.1. Natural Catastrophes **16**

1.1.1. Nat Cat coverage triangle: Insurers - Reinsurers - Government 16

1.1.2. Insured vs economic losses – Protection gap issue 18

1.2. Tornadoes **21**

1.2.1. Definition and features 22

1.2.2. Tornado genesis 22

1.2.3. Fujita scale 23

1.2.4. Facts and figures about tornadoes in the US 23

1.2.5. Increasing trend of tornado severity and its causes 24

1.2.6. Full set of tornado risk 25

1.3. Database: Tornadoes in the US over 1990 - 2008 ..... **27**

1.3.1. The variable *"t"* 27

1.3.2. The variable "Number of observations" 28

1.3.3. The variable "Aggregate cost in USD" 28

1.3.4. The variable "Cost of observations in USD" 28

PART I. Frequency-Severity Cat models under a Generalized Linear Model framework **30**

2. Frequency – Count Time Series **31**

2.1. A first overview of the count data **31**

2.1.1. Statistical description of the count data 32

2.1.2. Seasonal and Trend Decomposition using Loess 32

2.1.3. Monthly averages and variances 34

2.2. Theoretical preliminaries **36**

2.2.1. Generalized Linear Models (GLMs) 37

2.2.2. Modeling objectives regarding count data 38

2.2.3. Statistical learning methods and diagnostics 40

2.3. Selecting the model **44**

2.3.1. Classical Poisson (no trend/seasonality) 44

2.3.2. Quasi-Poisson (no trend/seasonality) 45

2.3.3. Poisson-trend 46

2.3.4. Poisson-seasonality 48

2.3.5. Poisson-trend-seasonality 50

2.3.6. Stationary COSINOR 52

2.3.7. Non-stationary COSINOR 53

2.4. The Frequency model **56**

3. Severity - Continuous distributions 58
3.1. A first overview of the claims data 58
3.1.1. Statistical description of the average cost data 59
3.1.2. Skewness feature and right-tailed distributions 61
3.2. Choosing the right-skewed distribution 64
3.2.1. Graphical comparisons 64
3.3. Theoretical preliminaries 71
3.4. Selecting the model 73
3.4.1. Classical Gamma (no trend/seasonality) 73
3.4.2. Gamma-trend 74
3.4.3. Gamma-seasonality 75
3.4.4. Gamma-seasonality (only months 4, 5, 8) 77
3.4.5. Gamma-trend-seasonality 79
3.5. The Severity model 81

Conclusion 84

Part II. Frequency-Severity Cat models with Multifractals 85

4. Multifractals – an innovative way 86
4.1. New features 86
4.2. Theoretical overview 88
4.2.1. Definitions and features 88
4.3. Applying multifractals to our schema 91
4.3.1. Frequency with multifractals 92
4.3.2. Severity with multifractals 95
4.3.3. Improvements of Frequency and Severity with multifractals 96
4.3.4. Two-dimensional Frequency-Severity with multifractals 97
4.3.5. Numerical Application 98

5. Derivative-reinsurance instruments – a numerical application 105
5.1. Traditional reinsurance treaties 106
5.1.1. Catastrophe per-event Excess of Loss - Cat XOL 107
5.2. Nontraditional reinsurance treaties 108
5.2.1. Insurance Linked Securities (ILS) context 108
5.2.2. Our proposal: “Asian Cat XOL” 109
5.2.3. Application of “Asian Cat XOL” 112

Conclusion, Limits, and Perspectives 117

Appendix A 123

Appendix B 125

References 127

List of tables

TABLE. 1 NATURAL CATASTROPHE LOSSES IN THE UNITED STATES, FIRST HALF 2016 19
TABLE. 2 NATURAL CATASTROPHE LOSSES IN THE UNITED STATES, 2006-2015.....19
TABLE. 3 FUJITA SCALE & ENHANCED FUJITA SCALE CATEGORIES (SOURCE: INSURANCE INSTITUTE).....23
TABLE. 4 EXTRACTION FROM SHELDUS TORNADO DATABASE 27
TABLE. 5 STATISTICAL SUMMARY OF THE TORNADO COUNT DATA 32
TABLE. 6 MONTHLY AVERAGES OF THE TORNADO COUNT DATA 34
TABLE. 7 MONTHLY VARIANCES OF THE TORNADO COUNT DATA 34
TABLE. 8 ESTIMATED PARAMETERS AND MONTHLY ESTIMATES FROM POISSON-SEASONALITY MODEL 48
TABLE. 9 ESTIMATED PARAMETERS FROM POISSON-TREND-SEASONALITY MODEL.....50
TABLE. 10 NON-STATIONARY COSINOR RESIDUAL STATISTICS OF THE OBSERVED TORNADO COUNT DATA 54
TABLE. 11.1 FREQUENCY MODEL COMPARISON.....56
TABLE. 11.2 ESTIMATED PARAMETERS OF FREQUENCY MODELS 57
TABLE. 12 SUMMARY STATISTICS OF THE TORNADO CLAIM DATA 59
TABLE. 13 AVERAGE COST OF THE TORNADO CLAIM DATA 60
TABLE. 14 ESTIMATED PARAMETERS AND MONTHLY ESTIMATES OF GAMMA-SEASONALITY MODEL 76
TABLE. 15 ESTIMATED PARAMETERS AND MONTHLY ESTIMATES OF GAMMA-SEASONALITY (ONLY MONTH4, MONTH5, AND MONTH8) MODEL 78
TABLE. 16 ESTIMATED PARAMETERS OF GAMMA-TREND-SEASONALITY MODEL.....79
TABLE. 17 SEVERITY MODEL COMPARISON 81
TABLE. 18 POISSON MULTIFRACTAL COMPARISON BASED ON THE NUMBER OF RANDOM FACTORS 100
TABLE. 19 OBSERVED MEAN OF TORNADO COUNT DATA VS POISSON MULTIFRACTAL SIMULATED MEAN 101
TABLE. 20 OBSERVED MEAN OF TORNADO CLAIM DATA VS POISSON MULTIFRACTAL SIMULATED MEAN 103
TABLE. 21 INSURER'S PAYOFF FOR AN ASIAN CAT XOL 112
TABLE. 22 MONTE CARLO PATHS: BASIC MODEL VS. MULTIFRACTAL.....114
TABLE. 23 PAYOFF: BASIC MODEL VS. MULTIFRACTAL.....114
TABLE. 24 PRICING RESULTS: BASIC MODEL VS. MULTIFRACTAL 114
TABLE 25. PROS AND CONS BY COMPARING BASIC MODEL & MULTIFRACTAL 116
TABLE 26. PROS AND CONS BY COMPARING VENDOR MODEL & MULTIFRACTAL.....121

List of figures

FIG. 1 BEAUTIFUL TORNADO NEAR SIMLA, COLORADO (SOURCE: PINTEREST PICTURES).....21
FIG. 2 TORNADO GENESIS (SOURCE: WIKIPEDIA).....22
FIG. 3 NATHAN WORLD MAP OF NATURAL HAZARDS 23
FIG. 4 INFLATION-ADJUSTED U.S. INSURED CAT LOSSES BY CAUSE OF LOSS, 1996-2015 (2015 $ BILLIONS).....24
FIG. 5 U.S. CONVECTIVE LOSS EVENTS, 1980-2015 (2015 $BILLIONS).....25
FIG. 6 MONTHLY NUMBER OF TORNADOES IN THE US.....31
FIG. 7 BOXPLOT OF THE TORNADO COUNT DATA.....32
FIG. 8 STL DECOMPOSITION OF THE TORNADO COUNT TIME SERIES 33
FIG. 9 GRAPH OF MONTHLY AVERAGES OF THE TORNADO COUNT DATA 35
FIG. 10 GRAPH OF MONTHLY VARIANCES OF THE TORNADO COUNT DATA 35
FIG. 11 FITTED VALUES OF POISSON-TREND MODEL VS OBSERVED COUNT DATA IN RED 47
FIG. 12 FITTED VALUES OF POISSON-SEASONALITY MODEL VS OBSERVED COUNT DATA IN RED.....49
FIG. 13 FITTED VALUES OF POISSON-TREND-SEASONALITY MODEL VS OBSERVED COUNT DATA IN RED 51
FIG. 14 COSINOR MONTHLY SEASONAL PATTERN OF THE TORNADO COUNT DATA.....52
FIG. 15 NON-STATIONARY COSINOR TREND AND SEASON 6-/12-CYCLES OF THE TORNADO COUNTS.....55
FIG. 16 MONTHLY AVERAGE COST OF THE TORNADO CLAIM DATA 59
FIG. 17 BOXPLOT OF THE TORNADO CLAIM DATA.....60
FIG. 18 MONTHLY AVERAGE COST OF THE TORNADO CLAIM DATA 61
FIG. 19 MONTHLY VARIANCES OF COST THE TORNADO CLAIM DATA 61
FIG. 20 SKEWNESS (SOURCE: WIKIPEDIA).....62
FIG. 21 HISTOGRAM COMPARISON OF THE CONTINUOUS DISTRIBUTIONS VS THE OBSERVED CLAIM DATA.....66
FIG. 22 OBSERVED TORNADO CLAIM DATA.....67
FIG. 23 FITTED GAMMA VS THE OBSERVED TORNADO CLAIM DATA.....68
FIG. 24 FITTED NORMAL VS THE OBSERVED TORNADO CLAIM DATA 68
FIG. 25 FITTED LOG-NORMAL VS THE OBSERVED TORNADO CLAIM DATA 69
FIG. 26 FITTED WEIBULL VS THE OBSERVED TORNADO CLAIM DATA 69
FIG. 27 FITTED VALUES OF GAMMA-TREND MODEL VS OBSERVED TORNADO CLAIM DATA IN RED.....75
FIG. 28 FITTED VALUES OF GAMMA-SEASONALITY MODEL VS OBSERVED TORNADO CLAIM DATA IN RED 77
FIG. 29 FITTED VALUES OF GAMMA-SEASONALITY (ONLY MONTH4, MONTH5, AND MONTH8) MODEL VS OBSERVED TORNADO CLAIM DATA IN RED.....78
FIG. 30 FITTED VALUES OF GAMMA-TREND-SEASONALITY MODEL VS THE OBSERVED TORNADO CLAIM DATA IN RED.....80
FIG. 31 FITTED VALUES OF BASIC FREQUENCY-SEVERITY MODEL VS THE OBSERVED TORNADO AGGREGATE COST DATA IN RED…….....85
FIG. 32 POISSON MULTIFRACTAL SIMULATED VALUES VS BACKTESTING OBSERVED SAMPLE 101
FIG. 33 GAMMA MULTIFRACTAL SIMULATED VALUES VS BACKTESTING OBSERVED SAMPLE103
FIG. 34 SCHEMA OF AN ASIAN CAT XOL TRANSACTIONS 111

Introduction

Catastrophe modeling (known as CAT modeling) is not considered as a brand new topic among researchers, given its numerous applications in different scientific fields. Engineers, meteorologists, or seismologists have been using it for the last decades to quantify damage severity, and a lot of work has been made in order to improve their models. Nowadays, one may find as many fitting applications in CAT modeling as one would like. In addition, exchanging theoretical results between fields is quite in common; a financer can find material for her/his model in what was formerly designed for a geophysical application [1]. The final goal is to have such models that could in addition predict with a certain pre-specified degree of precision the occurrence and the severity of catastrophic events. This is often realized by including probabilistic schemas in the case study, such as, for example, Bayesian approaches [2,3,4], Monte Carlo algorithms [4], or spatial prediction. Even though we are far from the "perfect" model, there will always be some place for improvement. Collaboration between researchers in knowledge-interchanging fields is necessary to design this kind of models.

Unsurprisingly, non-life insurance and reinsurance companies show great interest in the construction and development of CAT modeling. They need to evaluate natural catastrophe risks: when it is a non-life insurance company that sells the policy, because the underwritten contract mentions a related clause; or when one needs to know how much reinsurance must be purchased in a pool in the case of a reinsurance company. We can get the full understanding of the fact why insurers pay a lot of attention to when we consider the overall influence a natural catastrophe can have on a portfolio of, e.g., Home and Automobile coverage policies. The nearer a risky zone a policyholder lives, the higher the risk of claim declaration is (ex: earthquake in Italy, 2016). Recently, several reinsurers have also sold weather related derivatives to share risks. Risk measurement is up to actuaries and engineers working together in the corresponding departments.

To be more specific, Property & Casualty actuaries must necessarily take into account models including natural catastrophe (Nat Cat) risks. Indeed, they are used to modeling frequency (event arrivals) and severity (claim amounts) separately. We assume that event arrivals are independent from claim amounts. Once these two models are properly constructed, it suffices to multiply the average number of events resulted from the first with the average claim amount from the second, and so we obtain the expected aggregate loss amount. This is the classical way how actuaries resolve non-life insurance problems, when they calculate premiums to be paid by policyholders. In such a case, several goals rely on one's model:

- Counting

Whichever the natural catastrophe we are willing to model, from the start, a simple question intrigues us: "Which is the probability of occurrence of a natural catastrophe?". From there, we deduce the problem of counting these events and finding available statistical learning methods that can be used. One of the mathematical fields that studies such a phenomenon is Time Series, and more specifically Count Time Series. Simple discrete probability distributions such as Poisson or Binomial Negative are used to count

events. However, in our context, the problem is not as simple as it sounds, given the complex form of almost all databases in Nat Cat events. Different events of the same Nat Cat risk may have different return periods, which means that it will take different amounts of time for a similar event to occur in the future. If the average case is considered that of a Poisson distribution, denoted $N(\lambda)$, with parameter λ (lambda) also being the mean, then the mean equals the variance of the theoretical distribution, denoted $E(N) = VAR(N) = \lambda$. We are indeed far from the average case, with the variance usually being much higher than the average.

If the Poisson distribution is the starting point in modeling count data, many authors have used this distribution in their papers and modified its parameter in order to perform better fittings and/or a more accurate prediction. A drastic and necessary change is to include variability in forms of time trend and seasonality in the mean parameter [4,5]. Our objective in Chapter 2 is to construct an efficient count model that takes into account the features discussed above.

- Assessing losses

Catastrophes may have a huge impact there where they take place. Possessing a huge destructive capacity, they may affect a neighborhood, city, or even a whole country; human beings' lives may be in peril as well as their possessions such as vehicles, houses, and all kinds of buildings probably may get damaged or even destroyed. Having these possible scenarios in mind, the potential loss amount can sometimes reach billions of US dollars, leaving (re)insurers in the impossibility of paying immediately a lump sum covering the whole amount. The reserves found in the balance sheets of these two decision makers would be insufficient and state aid would be mandatory to resolve the situation.

This explains thus the importance of having a calibrated tool that would give a good estimate of the aggregate loss amount. In a brief description, aggregate losses may be of two kinds: attritional losses or extreme ones. It is particularly the second kind that must be well modeled. Classical distributions such as Normal, Gamma, Exponential, etc., do not quite catch extreme events while being fitted to observed data. Our objective in Chapter 3 is to construct a loss model which incorporates seasonality and trend of claims.

A last obstacle in modeling a frequency-severity couple is the independence assumption, which means the amount of claims is not explained by the number of catastrophic events at a given period of time. In other terms, there does not exist a certain structure showing some correlation between them. Nonetheless, the latest studies have been insisting on the fact that some pattern of dependence is obvious, but it differs in degree with the number of observations, the chosen sample, or the branch/branches of non-life insurance. It is shown statistically that the more tornadoes occur, the severer they are in terms of cost damage. One could ask if it is possible to find a hidden element affecting both the frequency and the severity of tornadoes. In these same studies, various proposals are made about how to make the dependence assumption part of non-life modeling. Technically, it is much more difficult to relax the independence assumption. The existing statistical learning methods would need huge changes, if a dependent structure were to be part of the model considered. For instance, we could mention the copula theory or

covariates based on the same random effects [4,5]. We deal with this problem in Chapter 4. Some correlation structure was added in the frequency-severity model in [4].

As we can understand from the last paragraphs, (re)insurers have a lot of difficulties when it comes to covering CAT-caused damages. These huge amounts not only overpass the reserves, but they also can often lead reinsurance companies to bankruptcy. In the last two decades, some innovative reinsurers have had brilliant ideas of merging the world of financial derivatives with theirs. It is what we call contingent derivatives that have come to their use. For example, swaps, contingent bonds [4,6], or weather derivatives have had a similar application in the world of reinsurance regarding CAT coverage. Based on the occurrence of CAT events, the possessor of a derivative will gain or lose from this contract. One then prefers such a financial instrument in order to share the risk of loss with another investor and avoid the risk of going bankrupt. Hence, a contingent derivative is an external layer that allows to enforce a CAT model.

To conclude this concise introduction, we are aware of tens of opportunities displayed in front of us to construct a CAT model. We also fully understand the importance of such a model, given its impact as it was discussed above. Following the method of a Frequency-Severity schema, adding new structures of seasonality, time trend, or volatility factors, and finally designing a derivative treaty over the estimated aggregated loss amounts, this is the path through which we will guide in this thesis. We do not claim that our way of modeling is the best compared to the already existing ones. We have somehow tried to improve the classical schema by adding this new original way. Our final aim is that it will serve to researchers, other actuaries or students in their projects.

1. Natural catastrophes in the US

1.1. Natural Catastrophes

As climate change issues are gaining more and more importance in world summits, adaption planning and/or global warming world program, it is not less for the way main decision-makers are trying to implement consequent figures arising from natural catastrophes in their policies worldwide. The specific insight that everyone currently has is the danger that a Nat Cat represents because of its direct impact on the population, business activities in the affected area and how this may quickly spread elsewhere. Government bodies, insurers and reinsurers have, in most cases, to respond to these extreme events according to their degree of involvement in a last of resort plan, underwritten insurance policy, or a reinsurance coverage by an international reinsurer, respectively. All three combined have a crucial role in the economy. One might ask how they work, together or separately, and what strengthens the necessity of their collaboration.

1.1.1. Nat Cat coverage triangle: Insurers - Reinsurers - Government

First of all, one must have a clear idea of how a Nat Cat risk is covered. It works similarly worldwide, but we will focus ourselves on the US model, for which we will be mentioning figures, organizations, or other historical data, as we model tornado risk, which occurs most severely in this area.

Let us start with the first actor from the private sector in this chain process. We are referring to the local insurance company. Without mentioning the hundreds of different insurance coverages, the two largest classes to be taken into account are property & casualty from one side and people lives from the other side. Depending on which risk an insurance company decides to cover by its policy proposed to the potential insured, several questions arise and need an answer, if one wants its coverage to function properly. Insurance companies are obliged to calculate their insurance premiums by applying actuarial-based practices, and this can be attained by using several discriminating factors. One must understand that a below-priced rate might cause disastrous consequences not only for the insurance company at first place, but also for the other role-players of the chain process. In order to not allow this to happen, one needs to access good quality data and refer to well-built mappings of the risk considered. Good quality data means possessing good measures of frequency and damage severity or other significant figures related to tornadoes. If a claim occurs, and if the pricing has been carefully undertaken, the insurer will encounter no problem covering the insured losses up to a maximum amount. In the case of Nat Cat, the coverage becomes quite expensive and may also suffocate an insurance company's capital and its ability to underwrite more policies. For decades, there has been a solution to this problem: ceding partial or whole portfolios of policies to the reinsurance companies.

Reinsurers help insurers by alleviating their charge of losses, and this is done in large parts, when a Nat Cat is covered. An insurer benefits from its relation with a reinsurer, because, while they cede, for example, a part of a portfolio, it doesn't have to cover that risk in its whole anymore and, as a result, less allocated capital and reserves liabilities are needed than before. Its balance sheet and Profit/Loss statement are improved and its underwriting capacity increases. Reinsurance companies operate in a wider geographical area, mostly internationally, and sometimes manage the same risk(s) in more than one continent at the same time. They have a better knowledge of such risks globally, and especially have the know-how of implementing catastrophe models. Via traditional and innovative reinsurance treaties, they are thus able to cover extra insured losses all around the world. But even these coverages work up to a limit too. It is up to governments to secure the rest of uninsured losses, or at least some part of it.

Government entities are considered to be an essential part of this chain process. They mostly intervene as a last resort decision-maker. In order not to collide with free market laws, a government will compensate for Nat Cat losses when "insurance/reinsurance protection is unavailable/unaffordable in the private market" [10]. The reimbursement schema is not as clear as it appears to be as, for example, in France. The French public reinsurer, i.e., Caisse Centrale de Réassurance (CCR), follows a specific pyramidal schema in case of an abnormal severity for certain natural catastrophes [7], and it covers damages in the very zones where insurance contracts have been underwritten. On the other hand, the U.S. have a federal government and every state may apply its own law regarding Nat Cat losses. So, one may identify two possible interventions: federally or on a state basis. Let us generally explain how these two ways work.

The federal government of the U.S. has constituted the so-called emergency funds and has created agencies such as the Federal Emergency Management Agency (FEMA) to compensate partly uninsured Nat Cat losses. To state a known extreme disaster, in the case of hurricane Katrina losses, FEMA has compensated at least $7bn through its "Individuals and Households Programme" alone [8]. There are also some programs like the National Flood Insurance Program (NFIP) that sell insurance coverage to property owners in return of a premium rate. On the other hand, each state may follow its own reimbursement policy. For instance, in California, the California Earthquake Authority (CEA) was created. It is a state-run insurance company that sold earthquake insurance coverage in the residential market through private insurance companies. This company and others alike are "ways of forcing insurance companies to bear, or to contribute to bearing risks they would otherwise shun" [9]. To sum up, there are two ways the federal or state government may handle the aftermath of a Nat Cat: by totally replacing the role of the (re)insurers or by acting as a last of resort helper.

Whether a government will fully or at parts pay for uninsured/under-insured losses, this is largely due to the pricing of coverages: underpriced if it doesn't exactly cover what it is stated in the contract, or overpriced because of a huge and uncontrollable potential damage. In both cases, one may guess that the homeowner's point of view reminds us of the "moral hazard" phenomenon: one will not buy an insurance policy, because one knows well that the government will anyway compensate one for losses caused by a Nat Cat. It is thus necessary and vital to boost the reliance towards (re)insurers, as they, based on actuarial studies, offer most of the time the fairest priced coverages. "Reinsurers also

monitor changes in weather patterns as part of their underwriting and risk evaluation process and use increasingly sophisticated catastrophe models to estimate expected losses from weather-related catastrophes" [10]. This is something a government is not able to do, given its lack of expertise on managing such risks. International (re)insurers are indeed fundamental to the US market, since they pay more or less than 60% of Nat Cat losses in the US.

It is now evident that, in order for this tripartite schema to maintain, a strong collaboration is needed from all parties. Lloyd's study on managing risks [10] came up with a list of 9 principles, which briefly tell the following:

Risk-based pricing in a healthy and private insurance market, where government intervention is minimized, is the fairest and most sustainable solution.

Specialist (re)insurers add value to the US Nat Cat market through additional capacity and expertise, based on good quality data and hazard mapping.

Governments and (re)insurers must respond to changing trends in the frequency and severity of losses, as they play an essential role in reducing the overall cost to the economy.

1.1.2. Insured vs economic losses – Protection gap issue

Before heading forward, let us first specify what the difference between economic losses and insured losses is, for this to have a significant impact on the schema discussed above. The Insurance Journal defines insured losses as "those that could arise from all exposures eligible for insurance coverage assuming standard limits and deductibles" [11]. By adding non-insurable losses, such as infrastructure damages or lost economic productivity, to insured losses, we get as a result an estimate of economic losses, follows the Journal [11]. Their estimate shows an insured loss standing at 44% of the overall economic losses for the US alone, and it goes up to 4 times higher for the global insured average annual loss. We can easily deduce the large gap it exists between insured and economic losses. In the same article, Sir Churney says that "economic loss estimates can be used to facilitate public risk financing and the development of regional resiliency plans to help societies better prepare for catastrophes and reduce the ultimate costs" [11]. His viewpoint converges well with Lloyd's principles.

The total economic losses from Nat Cat in the US since the early 90' have averaged tens of billions of dollars per year. The following table regroups the Nat Cat losses belonging to the first semester of 2016. According to the Property Claim Services (PCS®), the table here below, a Verisk Analytics® business, a catastrophe is defined as "an event that causes $25 million or more in insured property losses and affects a significant number of property/casualty (P/C) policyholders and insurers" [12].

As of July 12, 2016	Number of Events	Fatalities	Estimated Overall Losses (US $m)	Estimated Insured Losses (US $m)*
Severe Thunderstorm	21	25	11,600	8,500
Winter Storms & Cold Waves	8	55	2,300	1,500
Flood, Flash Flood	6	60	3,300	1,000
Earthquake & Geophysical	-	-	-	-
Tropical Cyclone	-	-	-	-
Wildfire, Heat Waves, & Drought (ongoing drought condition without loss estimation for the half year)	5	10	200	Minor losses
Totals	**40**	**150**	**17,400**	**11,000**

TABLE. 1 NATURAL CATASTROPHE LOSSES IN THE UNITED STATES, FIRST HALF 2016 [1]

A second table also based on perils is shown above. It is important to have a broader view of Nat Cat risks, such as a ten-year basis, for instance. This allows us to better reveal their average frequency and severity behavior from a national aspect.

As of February 2016	Fatalities	Estimated Overall Losses (US $bn)	Estimated Insured Losses (US $bn)*	10-year average insured losses (US $bn)*
Severe Thunderstorm	1,646	180	124	12.4
Winter Storms & Cold Waves	813	29	18	1.8
Flood, Flash Flood	333	33	6.5	0.65
Earthquake & Geophysical	5	1.5	0.4	0.04
Tropical Cyclone	425	128	65	6.5
Wildfire, Heat Waves, & Drought	511	77	28	2.8

TABLE. 2 NATURAL CATASTROPHE LOSSES IN THE UNITED STATES, 2006-2015 [2]

Whichever the Nat Cat considered is, the corresponding economic losses are aggravating from year to year. Several factors mostly explain these increasing key figures. Only the US population in Nat Cat-prone areas has grown by 70% during the last 50 years. On the

[1] Source: © 2016 Munich Re, NatCatSERVICE; Property Claim Services (PCS®) *, a Verisk Analytics® business. As of July 2016.

[2] Source: © 2016 Munich Re, NatCatSERVICE; Property Claim Services (PCS®) *, a Verisk Analytics® business. As of March 2016.

other hand, the households' wealth mainly composed of property has grown too. According to a 2008 study by AIR Worldwide, from December 2004 to December 2007, "the insured value of properties in coastal areas of the United States continued to grow at a compound annual growth rate of just over 7%" [13]. Finally, a much higher business activity related to a significant economic development during the same period makes the US economy under Nat Cat impact weaker than before. These changes generate stronger loss severity, and extreme events become consequently even more extreme. However, "scientists are attributing a significant portion of the increases in storms, temperature extremes, droughts, wild fires and floods to climate change", says Lloyd's report [10].

These figures show, once more, that governments, insurers and reinsurers working together is the only way to handle these extreme risks.

In the rest of this chapter, we will particularly tackle the issue of one of these Nat Cat from the category of Severe Thunderstorms, namely, the tornadoes. All the models in this thesis has been constructed based on the observed data of tornadoes in the US; a thorough overview of this risk appears to be fundamental, if one wants one's model to incorporate the features of tornado risk.

1.2. Tornadoes

As it was mentioned before, to understand the danger a risk poses and how much severe it can become, it is necessary to explore the genesis of such risk, i.e., redefining, classifying, and measuring it in terms of potential losses. From a general point of view, the scientific community agrees that one can define risk as the "product" of hazard and its adverse consequences, say, the exposure and vulnerability. Also, an event related to that risk, such as a tornado, is termed a catastrophe only if people are harmed and/or their positions damaged [12]. We went through several studies such as Insurance Information Institute's official website, Munich Re's [12], and Lloyd's report [10], and realized at first that a tornado event is categorized in the class of severe storms, and is mainly formed on a thunderstorm basis. For more clarity, a towering cloud with strong updrafts and downdrafts is called a thunderstorm. According to the National Weather Service, they are called "severe" if "they cause the formation of a tornado" [12].

Some natural questions arise at this stage: What is a tornado? How does it occur? Do we know the causes behind the occurrence of such a storm event? And finally, what do figures tell us about the impact of tornadoes in the US? To accurately answer these questions, not only did we base our research on the former studies, but we also referred to other scientific facts about tornadoes. An extraction of the data used for modeling is also given and explicated in the end of this chapter.

Fig. 1 Beautiful tornado near Simla, Colorado [3]

[3] Source: Pinterest pictures

1.2.1. Definition and features

Let us first define this event. According to the National Oceanic and Atmospheric Administration & Columbia Encyclopedia, a tornado begins with "a dark, funnel-shaped cloud which extends from a thunderstorm", and basically contains "a violently rotating column of air that will eventually come into contact with the ground" [14].

Briefly say, a tornado is generated from a thunderstorm, has the form of a violently rotating column of air, and hits the earth. Its column of air can have a diameter varying from a few meters to more than 700 m. A distinguished feature of a tornado is the speed of its movement. It can reach velocities up to 480 km per hour and can cross paths starting at a few miles to more than a hundred. Its known movement generally is the Northeast direction, even though it can move towards any of the other possible directions. The atmospheric conditions typically required for the formation of a tornado include "great thermal instability, high humidity, and the convergence of warm, moist air at low levels with cooler, drier air aloft" [14]. Finally, it is usually accompanied by thunder, lightning, heavy rain, and a loud "freight train" noise [14].

1.2.2. Tornado genesis

Another interesting fact is related to the tornado-genesis, which is the process of the tornado formation. Before stating the different ways of formation, it is interesting to let know that there are still some unknown climatic aspects of the tornado formation undergoing scientific study. Indeed, one of Multifractal model's core assumptions in Chapter 4 was made given the mystery of these, say, climatic factors.

To continue with the ways how tornadoes are formed, one can distinguish 3 of them: super-cellular, land-spouts, and waterspouts. We are mainly interested in super-cellular tornadoes, which occur on a terrestrial area. The birth of this kind is described as follows: "First, the rotating cloud base lowers. This lowering becomes a funnel, which continues descending while winds build near the surface, kicking up dust and other debris. Finally, the visible funnel extends to the ground, and the tornado begins causing major damage" [15]. Consecutive pictures of a tornado formation are given below:

Fig. 2 Tornado genesis [4]

[4] source: Wikipedia

1.2.3. Fujita scale

The Fujita scale, a "National Weather Service" product, measures the intensity of a tornado given the level of damage they cause [16]. We distinguish 6 categories (see table below) labeled with the corresponding degree of damage severity. Around 50% of tornado cases fall into the first category. With a speed of 73-112 mph, they can "overturn automobiles and mobile homes, rip off the roofs of houses, and uproot trees" [17]. The F5 category alone regroups approximately 1% of tornado cases. When the wind speed exceeds 261 mph, they are capable of "lifting houses off their foundations and hurling them considerable distances" [17].

		Original F scale (1)	Enhanced F scale (2)
Category	Damage	Wind speed (mph)	3-second gust (mph)
F-0	Light	40-72	65-85
F-1	Moderate	73-112	86-110
F-2	Considerable	113-157	111-135
F-3	Severe	158-207	136-165
F-4	Devastating	208-260	166-200
F-5	Incredible	261-318	Over 200

TABLE. 3 FUJITA SCALE & ENHANCED FUJITA SCALE CATEGORIES (SOURCE: INSURANCE INSTITUTE)

As we can see, there is also an Enhanced version. The difference with the former is that it uses 28 damage indicators consisting of buildings, structures, and trees when it determines the EF-scale, which thus drastically improves the former version. Each one of these indicators has a standard description of the typical construction for that category, and ranges from trees to shopping malls.

1.2.4. Facts and figures about tornadoes in the US

A closer look at the map [18] furnished here below highlights the importance of tornado risk, as being the most frequent in the US zone.

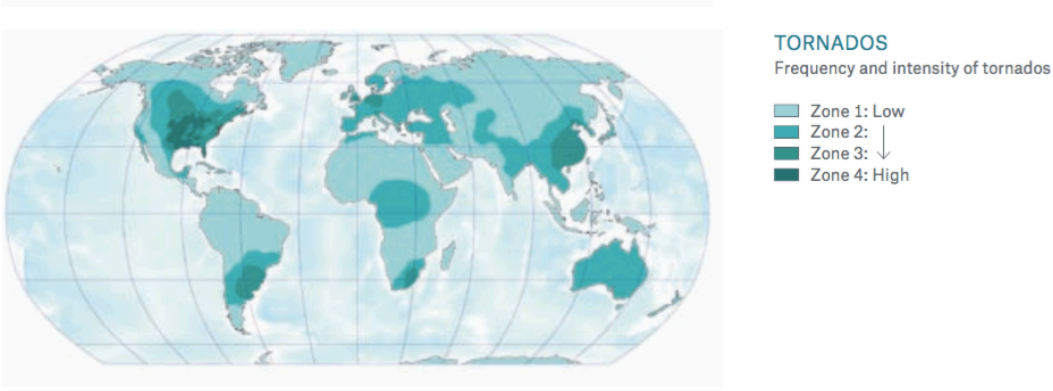

Fig. 3 Nathan world map of natural hazards

More specifically, the Continental US, or the Central and the Southern parts, including the Ohio valley and the Gulf states, represent the areas where the atmospheric conditions required for a tornado formation are met. The tornado-prone area is the so-called Tornado Alley, extending at some point from the Rocky Mountains to the Appalachians, where the warm and humid air coming from the Gulf of Mexico collides with the cool and dry air from the Rockies and Canada [19].

The highest frequencies occur through April-June. Even though the frequency peaks are found in May and June observations, April appears indeed to be the most fatal month in deaths. One of the worst cases was recorded on April 27, 2011, when 137 reported tornadoes took place in six of the Southern states by killing around 300 people. It is measured that 60 people on average die annually from flying and falling debris [19].

1.2.5. Increasing trend of tornado severity and its causes

"The number of weather-related loss events in North America for the past three decades has nearly quintupled", says Munich Re report [12]. Inflation-adjusted property insurance market losses follow after. "Between 1980 and 2011, 43% of insured property windstorm losses in the United States in 2011 values were caused by severe thunderstorms" [12]. The highest record of damages caused by thunderstorms was reported in 2011, and estimated to be around US$ 26bn. A data example with proportions is shown below. Tornado risk leads all the other perils.

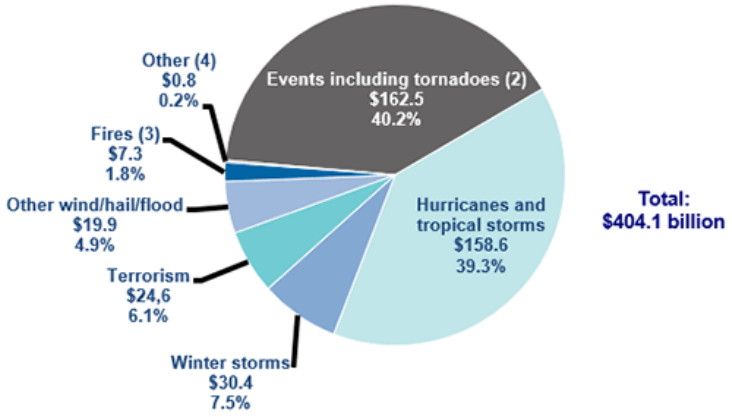

Fig. 4 Inflation-Adjusted U.S. Insured Cat Losses by Cause of Loss, 1996-2015 (2015 $ billions)[5]

Studies clearly show an upward time trend for the selected period. If there is one point to which more than 90% of scientists agree is the fact that anthropogenic climate change hugely contributes to this trend; scientific findings are supporting the cause of global

[5] Source: Property Claim Services (PCS®), a Verisk Analytics® business

warming as being behind more frequent and more intense weather extremes [12,20]. Climate change, accompanied by certain socio-economic factors such as population growth, spread of urbanization, and a higher wealth of North-Americans, strongly impacts the increase of economic losses. As these factors are likely to become even more significant in their direct consequences, there is little doubt that losses also are expected to follow the same path. The graph below perfectly shows this upward trend and the existing protection gap.

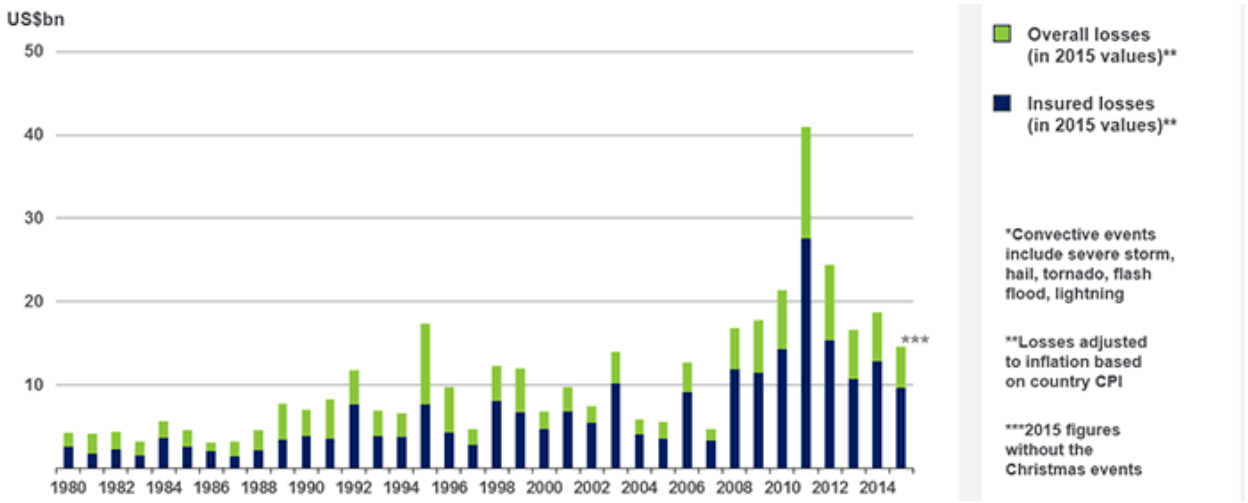

Fig. 5 U.S. Convective Loss Events, 1980-2015 (2015 $billions)[6]

Remark: Unless stated otherwise, time trend and trend will be used interchangeably.

1.2.6. Full set of tornado risk

One would finally ask how this risk is considered by the (re)insurance industry. Tornado risk is usually covered under insurance policies commercialized by property business lines. For example, in our case, the windstorm coverage or the all-risks coverage accept as a charge some, but rarely all, of tornado damages. In those areas exposed to severe thunderstorms, a specific deductible may also be applied.

Regarding the evaluation of tornado risk, while frequency and severity are said to be first-hand hazard components, we can also take into account indirect components such as:

- Exposure, say, the total amount of real estate, building, residential and commercial values present in the tornado-prone area
- Vulnerability, say the potential damages

[6] Source: © 2016 Munich Re, NatCatSERVICE; Property Claim Services (PCS®)*, a Verisk Analytics® business. As of January 2016.

One need to know that the vulnerability of a zone in particular is less than or equal to the exposure of that same zone. The higher the vulnerability of a zone is, greater risk is generated when causing damage losses.

To sum up, the structure of hazard, put together with exposure and vulnerability, leads to a full evaluation of tornado risk. [12] Now that we have a complete insight of such risk, we may present the tornado database that we will use in our study, and show how we can recognize the hazard components.

1.3. Database: Tornadoes in the US over 1990 - 2008

Tornadoes represent a very frequent issue in the US. Being a meteorological phenomenon, a great variety of climatic factors are found in its causes. Even though weather-related predictive programmed tools exist in the market, we still do not have full knowledge of how a tornado happens. We need thus a sufficient amount of data to output accurate results.

The data about monthly tornado arrivals and insured costs were retrieved on "SHELDUS" [21]. The way how SHELDUS database is constituted is explained above. From this database, we possess a sample of 228 monthly tornado data through the years 1990 - 2008 showing the following: the number of month when occurred, the number of tornadoes, and the aggregate insurance claims amount. The average insurance claims amount was also calculated for modeling purposes.

We further note that count, arrival, and number of tornadoes will be used interchangeably. Insured/tornado cost, claim/cost severity, and insured/insurance claims will be used interchangeably as well.

Hereby stands an extract from the database:

t	Number of observations	Cost of observations in USD	Aggregate cost in USD
1	7	389,115.15	2,723,806.08
2	27	391,298.74	10,565,066.01
3	61	787,508.76	48,038,034.30
4	16	1,428,966.44	22,863,463.07
5	48	383,465.12	18,406,325.90
6	104	858,067.69	89,239,040.26
7	13	82,539.58	1,073,014.54
…	…	…	…

TABLE. 4 EXTRACTION FROM SHELDUS TORNADO DATABASE

1.3.1. The variable *"t"*

It is an indicator of time showing in which step of the time period 1990 - 2008 we are. For example, the number 1 stands for January 1990, 2 for February 1990…, 228 for December 2008.

1.3.2. The variable "Number of observations"

Let us denote by N the number of tornadoes that have occurred till month t, say, $N(t)$. It is necessary to specify that we are looking for the number of tornadoes that have occurred during each month. Thus, every month, we are interested in $N(t) - N(t-1)$, or the number of tornadoes occurred during month t only. This quantity appears in the 2nd column of the table shown above. For example, 48 tornadoes occurred during May 1990. The number of tornadoes per month in SHELDUS is calculated according to Fujita scale measurements, from scale F0 to F5.

This variable represents the first consequence of tornado risk, the hazard, which will later result in a probability of occurrence.

1.3.3. The variable "Aggregate cost in USD"

Let us denote by S the aggregate insurance claims amount accumulated till month t, say $S(t)$. We are interested in the aggregate claims amount during month t only, which is $S(t) - S(t-1)$. We can find this quantity in the 4th column of Table 4. For example, we had an overall insurance claims cost of $18,406,325.90 in May 1990. SHELDUS database gives thus the total direct cost per each month during the 1990-2008 period considered in this thesis. This amount represents property and crop losses; losses from business interruption are not included. In order to avoid inflation bias, the amounts are represented here in an "As-If" basis, which means in 2008 US dollars. These amounts do incorporate, anyway, the exposure bias: we have aggregate amounts, so the information regarding the affected location (county and state) is lost here. This won't be a problem, since we work on a national basis.

This variable represents the second and forth consequences of tornado risk, the severity and direct losses from property (re)insurance coverages.

1.3.4. The variable "Cost of observations in USD"

It represents the average of insurance claims amount regarding the month t. It is found in the 3rd column of the data table. For example, we had an average cost of $383,465.12 per tornado in May 1990. One would calculate it as follows:

$$\text{“}Cost\ of\ observations\ in\ USD\ \$\text{”} = \frac{\text{“}Aggregate\ cost\ in\ USD\ \$\text{”}}{\text{“}Number\ of\ observations\text{”}}$$

The table resumes thus the features of tornado risk. The variable *"t"* will serve as the basic time explanatory variable throughout this thesis. Regarding the frequency model in Chapter 2, the variable *"Number of observations"* is taken into account. In Chapter 3, we consider the variable *"Cost of observations in USD"*.

With the database being set and tornado risk fully described, we are now prepared to study each one of tornado risk components separately and learn how to model them with the Frequency-Severity model in Chapter 2 and 3, or the Multifractal model in Chapter 4.

PART I. Frequency-Severity Cat models under a Generalized Linear Model framework

2. Frequency – Count Time Series

In this first chapter, our main work will consist in modeling one constitutive element of hazard regarding tornado risk. For the record, in Chapter 1, a thorough overview of tornado risk, and more generally Nat Cat risks, was presented. It was specified that hazard puts together frequency and severity of a weather extreme, which are the first components to be modeled. We will obviously start by undertaking a careful study on the frequency of tornado risk. Later on, several models will be compared and only one of them will be kept for further examination vis-à-vis the Multifractal model in Chapters 4 and 5. Above all, it is absolutely necessary to highlight the importance of count process when dealing with natural catastrophes. Since we are handling a weather-related event, climatic factors behind it become a major part of this study. Remember that we emphasized the rigorousness of climate change and the visible upward trend in the last three decades. Adding to that seasonality of tornado arrivals, one might construct a reliable model. Consequently, one must not only predict when the next tornado occurs, but should also implement some sort of trend and a seasonal structure, as these both enforce the quality of prediction.

2.1. A first overview of the count data

We refer ourselves to the variable "*Number of observations*", denoted *N* in the description of the tornado database. The number of tornadoes occurred in the US is given per each month during the 19-year period. It should be specified again for accuracy that the covered area are the US. Let us now visualize the count data with the help of the graph given below.

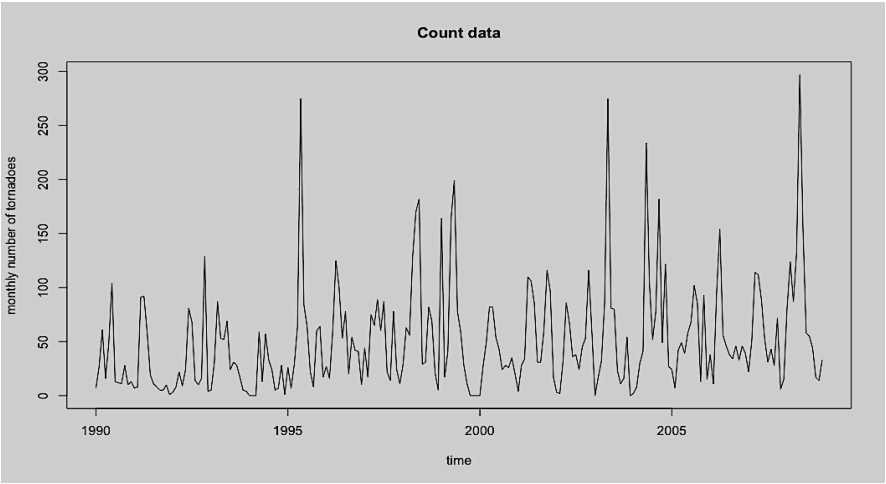

Fig. 6 Monthly number of tornadoes in the US

2.1.1. Statistical description of the count data

Some statistical characteristics of the count data are made available through the following summarized table. R and SAS software are used to output all the following results. Let us first pick the monthly average of the number of tornadoes in the US. In our case, approximately 51 tornadoes occur per month during the chosen period. The graph already indicates that there are some extreme numbers of tornadoes occurred in one month, with the highest peak being of 297 tornadoes in 2009. By the use of a statistical tool, the boxplot shown below, we can also deduce the severe skewness on the right. In other terms, there are great deviations above the monthly average of the number of tornadoes. As a result, we should expect a very strong variability in the count data.

Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
0	16	35.5	50.71	69	297

TABLE. 5 STATISTICAL SUMMARY OF THE TORNADO COUNT DATA

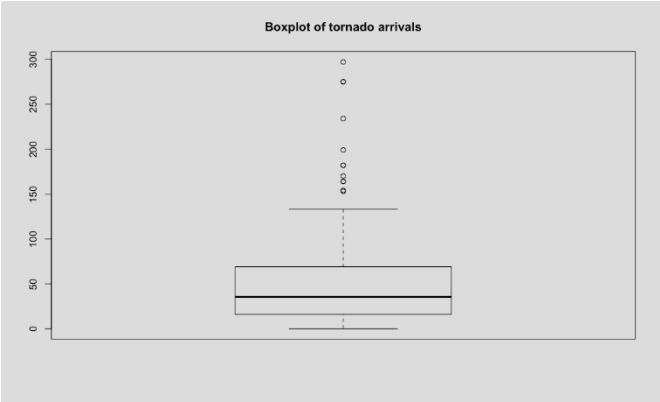

Fig. 7 Boxplot of the tornado count data

2.1.2. Seasonal and Trend Decomposition using Loess

Some other statistical tools could make explicit some other interesting features of the count data. Loess decomposition of a time series is one of them. In mathematical sciences, count data are referred to as a time series. A time series usually has three components: general trend, seasonality and the error term. In our case, a fourth component, representing jumps, ought to be added. More specifically:

General trend shows the general direction of change in count data.
Seasonality represents the seasonal structure of data per year.
The error term is the random component of the time series, which is not explained by the other components, but affects them in the forthcoming period.
Jumps are given by peaks of the number of tornadoes here.

The R-package in question applies the "Seasonal Trend Decomposition" using Loess (STL), an algorithm that was developed to help divide up a time series into the three components described above. The Seasonal Trend decomposition was developed by R. Cleveland, W. Cleveland, J. McRae and I. Terpenning in the Journal of Official Statistics in 1990 [22]. It is said to be a versatile and robust method. Two frameworks are needed to put the algorithm into practice, a seasonal structure and trend. Loess, on the other side, is a method for estimating nonlinear relationships. The two methods combined, by the help of the algorithm behind the R-package, generate the following output:

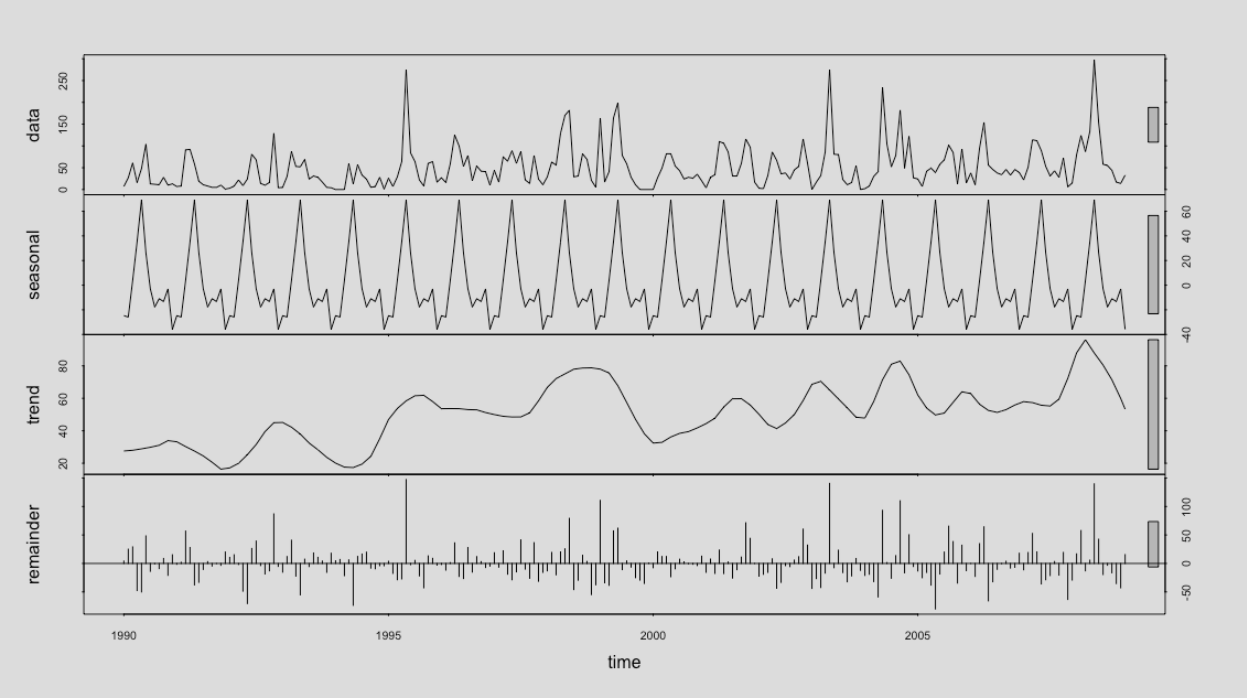

Fig. 8 STL Decomposition of the tornado count time series

The first part of the graph represents the data. The second one shows the seasonal component. We realize here that it certainly exists a seasonal behavior in the count time series. The seasonal form becomes severer with the time. The third part is the trend component. There is a slight upward moving trend in the data, which goes along with Chapter 1 statements. The final part shows the error component. From the remainder graph, we can tell that the count data is quite volatile. As far as we are concerned about the jump component, another kind of analysis will be used in Chapter 4.

These initial conclusions need to be verified with more sophisticated analyses. Otherwise, they will be used as empirical arguments before we fit models in this Chapter and the next two ones. On the other hand, a more concrete analysis strengthening some of these decomposition method results will be shown in the following section.

2.1.3. Monthly averages and variances

To have a closer understanding of seasonality and high volatility concepts, there is no better way than referring to the observed data themselves. For this purpose, we found it necessary to furnish some statistical figures: the empirical monthly averages (Table 6) and variances (Table 7). For instance, the monthly average/variance in May is calculated with May observations only. Graphs (Figures 9 and 10) help us visualize each statistic.

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
25.53	24.63	54.53	85.74	119.68	76.74	47.63	32.95	39.84	37.74	48.16	15.37

TABLE. 6 MONTHLY AVERAGES OF THE TORNADO COUNT DATA

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
1542	839	899	1920	8485	1512	516	532	1813	938	1848	240

TABLE. 7 MONTHLY VARIANCES OF THE TORNADO COUNT DATA

In Table 6, we can confirm well the fact that the "tornado season" is known to be as spring, because a majority of them occur from April through June, with May being the peak. In Table 7, the variance in May is relatively very high. This makes us think of an undervaluation of the other monthly variances. When one has a look at Figure 9, one can see how the monthly average varies.

An annual, as well as a semi-annual seasonal pattern may be considered for further modeling. In the second case, a double-seasonal pattern appears. We are far from the simple case when the mean equally varies for each monthly path. The variability of the monthly average here implicitly incorporates somehow the variability of the observed data too.

Figure 10 is another valid representation of the observed data variability. It makes it possible for us to distinguish a similar way of how the monthly variance moves from one month to another. If one compares between Figures 9 and 10, one can think of the possibility that a certain correlation structure could be designed. Some studies [24] have already taken into account this possibility. In this chapter, we have also put in place an over-dispersion method in Section 2.3.2, which allows us to have a more specific understanding of this structure.

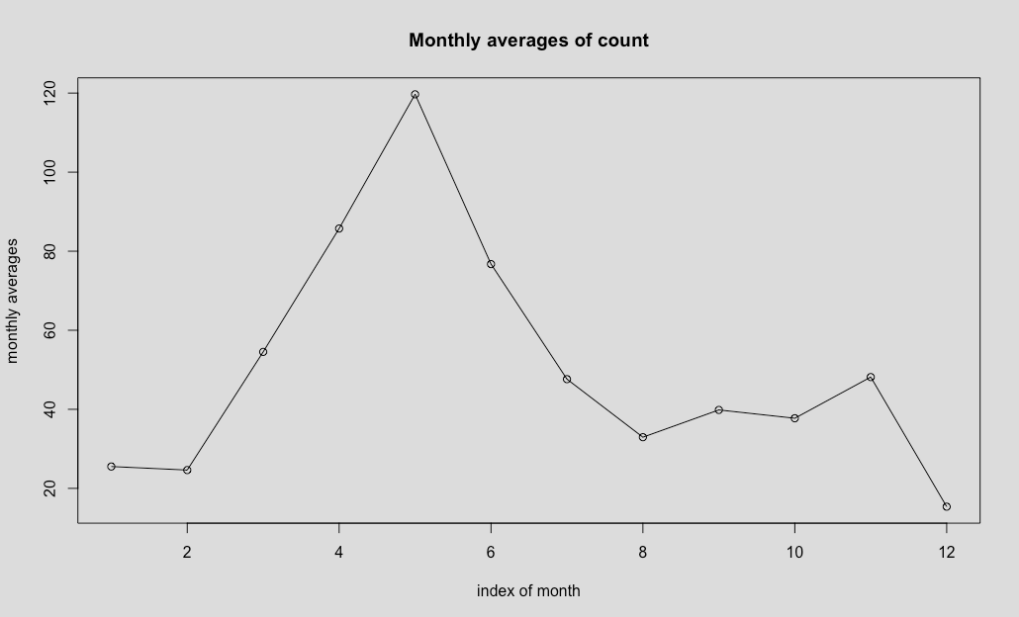

Fig. 9 Graph of monthly averages of the tornado count data

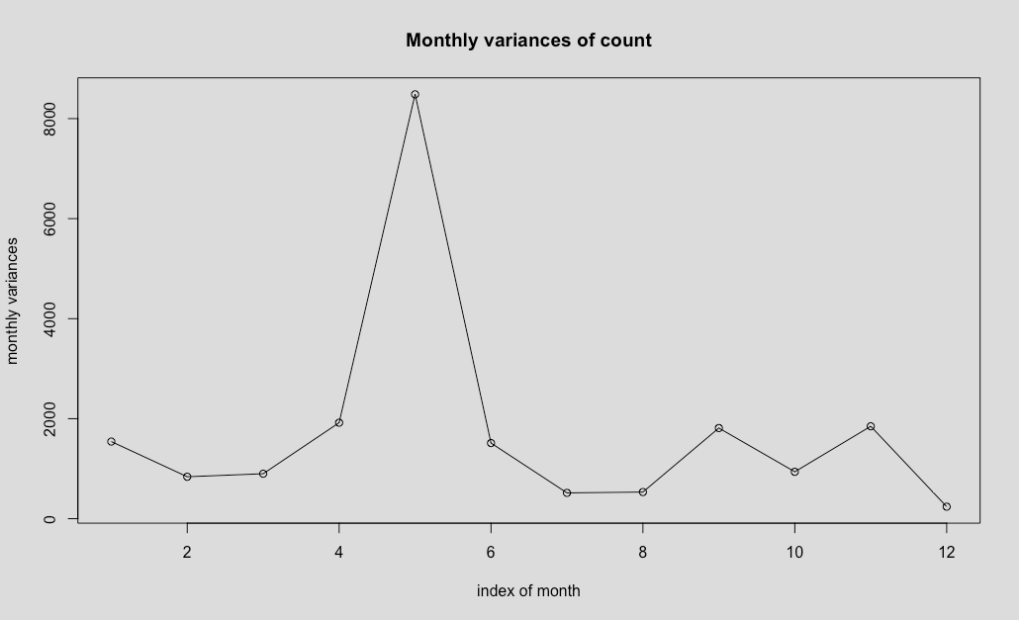

Fig. 10 Graph of monthly variances of the tornado count data

Both empirically and graphically, we showed some quite interesting results about the tornado count data that will serve us for the rest of this thesis. One also needs to clarify the theory of frequency model testing, before one performs such methods. It will be the purpose of the following section.

2.2. Theoretical preliminaries

A theoretical review is detailed in this section for several reasons. We want the model framework to be clearly understood. Some other statistical learning methods and decision criteria are also essential when one should choose a frequency model for the count data. For this purpose, we will define some useful concepts and give a brief description per each method and/or criterion accompanied by the corresponding formulae.

While dealing with statistical inference, information known up to time T is taken into consideration. One clever way of performing this methodology is via Partial Likelihood. Let $\{Y_t\}$, $t = 1, 2, \dots, N$, be a time series having the following joint distribution $f_{\boldsymbol{\theta}}(y_1, y_2, \dots, y_N)$, where the parameter set is given by the vector $\boldsymbol{\theta}$. Partial Likelihood becomes a smart tool especially when the time series is observed at the same time with some "random time dependent covariates", which applies very well to tornado data nature (resumed from section 1.1 of [23]). The definition is given hereby.

Partial Likelihood, Def. 1.1 of [23]: Let $\mathcal{F}_t$, $t = 0, 1, \dots$ be an increasing sequence of σ-fields, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \dots$, and let $Y_1, Y_2, \dots$ be a sequence of random variables on some common probability space such that Y_t is $\mathcal{F}_t$-measurable. Denote the density of Y_t, given $\mathcal{F}_{t-1}$, by $f_t(y_t;\ \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^p$ is a fixed parameter. The partial likelihood (PL) function relative to $\boldsymbol{\theta}$, $\mathcal{F}_t$, and the data $Y_1, Y_2, \dots, Y_N$, is given by the product:

$$PL(\boldsymbol{\theta}; y_1, y_2, \dots, y_N) = \prod_{t=1}^{N} f_t(y_t;\ \boldsymbol{\theta}).$$

In the partial likelihood, we take into account only the information that is available up to the current time T. This is why we call it "sequential conditional inference" [23]. $\boldsymbol{\theta}$ represents the maximum partial likelihood estimator - MPLE. Theoretically, this estimator is consistent and efficient.

Let us go further with the inference framework. For $\{Y_t\}$, $t = 1, 2, \dots, N$, being the time series we are interested in, which we call the response, let then $Z_{t-1} = \left(Z_{(t-1)1}, \dots, Z_{(t-1)p}\right)'$ be the vector containing the covariates, also called past explanatory variables. If we denote by $\mathcal{F}_{t-1}$ the following σ-field: $\mathcal{F}_{t-1} = \sigma\{Y_{t-1}, Y_{t-2}, Z_{t-1}, Z_{t-2}, \dots\}$, then, our problem is to relate the response to the covariates of Z_{t-1}. This problem can be resumed mathematically as follows:

$$\mu_t = \mathbb{E}[Y_t | \mathcal{F}_{t-1}],$$

where μ_t is the conditional expectation of the response given the past.

2.2.1. Generalized Linear Models (GLMs)

A classical way of dealing with parameter estimation consists in using the Linear Regression inference methods. In our case, this is in particular not interesting, because they may generate negative values of the Poisson or Gamma mean, random processes of interest in the later sections. Even though Linear Models have a simple and easily interpretable inference method, their quality of prediction often is weak compared to other prediction methods. Generalized Linear Models (GLMs) are a solution to this kind of obstacles. The general ideas ruling GLMs, Exponential Families and link functions, can be applied to time series as well [23].

Time series following GLMs, Def. 1.6 of [23]: The two following stipulations reign their behavior:

Random Component. The conditional distribution of the response given the past belongs to the exponential family of distributions in *natural* or *canonical* form. That is, for $t = 1, 2, \dots, N$,

$$f(y_t;\ \theta_t, \phi | \mathcal{F}_{t-1}) = exp\left\{\frac{y_t\theta_t - b(\theta_t)}{\alpha_t(\phi)} + c(y_t;\phi)\right\}.$$

The parametric function $\alpha_t(\phi)$ is of the form ϕ/ω_t, where ϕ is a *dispersion* parameter and ω_t is a known parameter referred to as *weight* or *prior weight*. The parameter θ_t is called the *natural* parameter of the distribution.

Systematic Component. For $t = 1, 2, \dots, N$, there is a monotone function $g(\cdot)$ such that:

$$g(\mu_t) = \eta_t = \sum_{j=1}^{p} \beta_j\, Z_{(t-1)j} = Z'_{t-1}\beta.$$

The function $g(\cdot)$ is called the *link function* while η_t is referred to as the linear predictor of the model.

One can easily find the following relationships in Section 1.2 of [23]:

Between the conditional mean and the natural parameter θ_t:

$$\mu_t = \mathbb{E}[Y_t|\mathcal{F}_{t-1}] = b'(\theta_t) \Leftrightarrow \theta_t = (b')^{-1}(\mu_t)$$

Between the conditional variance and the natural parameter θ_t:

$$Var[Y_t|\mathcal{F}_{t-1}] = \alpha_t(\phi)b''(\theta_t) = \alpha_t(\phi)V(\mu_t).$$

The function $V(\mu_t) = b''(\theta_t)$ is called the *variance function*, depending on μ_t.

Finally, $g(\mu_t) = \theta_t(\mu_t) = \eta_t = Z'_{t-1}\beta$ is said to be the *canonical link* function.

Another representation would be, for $g = \mu^{-1} = (b')^{-1}$ being the explicit form of the canonical link function,

$$\theta_t = (b')^{-1}\big(g^{-1}(\eta_t)\big) = \mu^{-1}\big(g^{-1}(\eta_t)\big)$$

where $\mu^{-1}(\cdot) \equiv (b')^{-1}(\cdot)$, and $\frac{\partial\theta}{\partial\mu} = 1/V(\mu_t)$.

Why should we use GLMs when one might simply apply Linear Regression analysis? There are some fundamental arguments to convince one for the beneficial aspects of GLMs. Linear Models attempt to fit a model to the mean response "μ" of some observed variable Y in the form of a linear predictor "η". GLMs are an extension to this approach. GLMs allow for a non-linear function "g" of the mean to be modeled in terms of a linear predictor. They also allow the error distribution to be a member of the exponential dispersion (ED) family, thus greatly broadening the set of distributions that can be fit to count data. GLMs finally have the special feature to include a mean-variance relation in the model, which is inherent in the exponential dispersion models density structure [24]. Thus, "in modeling the mean through a GLM, we are also indirectly modeling the variance" [24].

2.2.2. Modeling objectives regarding count data

Our 1st goal is to model the mean response (average number of tornadoes or average claim amount) conditional on a given set of covariates Z_{t-1}. In the analysis of insurance data, covariates are often referred to as "rating variables" and are used to characterize the risk ensued by some insured individual. Furthermore, the GLM framework, as discussed in McCullagh and Nelder, 1989, assumes that the random variables Y_i for $i = 1, \dots, N$ are independent, and that they follow exponential dispersion distributions, say, $Y_i \sim ED\ (\mu_i, \varphi)$: their means vary with each observation, while the dispersion is assumed to be the same for all the observations, but is considered as unknown.

$$\mu_t = \mathbb{E}[Y_t|\sigma(Z_{t-1})] = g^{-1}(\eta_t) = g^{-1}(Z'_{t-1}\beta)$$

2.2.2.1. Poisson distribution in "Time Series following GLMs"

The count time series here, namely, the tornado arrivals data, will be modeled through the Poisson distribution. "Time Series following GLMs" 's framework leads us to the following:

$$f(y_t;\ \theta_t, \phi|\mathcal{F}_{t-1}) = \exp\left((y_t \log\mu_t - \mu_t) - \log y_t!\right),\ t = 1, 2, \dots, N.$$

So that we have the following theoretical assumptions (Page 9 of [23]):

$$\mathbb{E}[Y|\mathcal{F}_{t-1}] = \mu_t,\ b(\theta_t) = \mu_t = \exp(\theta_t),\ V(\mu_t) = \mu_t,\ \phi = 1, \text{and } \omega_t = 1$$

The canonical link is:

$$g(\mu_t) = \theta_t(\mu_t) = \log \mu_t = \eta_t = Z'_{t-1}\beta.$$

In order to understand the pattern of a canonical link, let us give a simplified example (Page 9 of [23]): For $Z_{t-1} = (1, X_t, Y_{t-1})'$,

$$\log \mu_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1},$$

with $\{X_t\}$ standing for some covariate process, such as a possible trend, or a possible seasonal component.

The 2nd goal of the GLM approach is to estimate the regression parameters β to ultimately predict the response variable Y. In order to maximize the Partial Likelihood, $PL(\beta) = \prod_{t=1}^{N} f(y_t;\ \theta_t, \phi|\mathcal{F}_{t-1})$, we calculate the log-PL, denoted by $l(\beta)$. The maximization problem is given by:

$$\max_{\beta} l(\beta) = \max_{\beta} \sum_{t=1}^{N} \log f(y_t;\ \theta_t, \phi|\mathcal{F}_{t-1}),$$

and the solution is given by $\hat{\beta}$, also called the maximum partial likelihood estimator - MPLE - of β in Definition 1.23 of [23]. This solution is found by the Fisher scoring method. There is also a Fisher information matrix, given the information $\mathcal{F}_{t-1}$, denoted by $G_N(\beta)$. For more details over the maximization procedure and the conditional Fisher matrix, refer to Section 1.3 of [23].

2.2.2.2. Poisson Regression

Let $\{Y_t\}$ be a count time series assumed to follow the Poisson distribution, and $\{Z_{t-1}\}$ the corresponding covariate process, for $t = 1, \dots, N$. Then,

$$f(y_t|Z_{t-1}) = \exp\left((y_t \log \mu_t - \mu_t) - \log y_t!\right),\ y_{t=0,1,2,\dots,N}$$

The conditional mean and variance are equal here, $\mu_t = \sigma^2$. Assuming the canonical link,

$$\log \mu_t = \eta_t = Z'_{t-1}\beta \leftrightarrow \mu_t(\beta) = \exp\left(Z'_{t-1}\beta\right),$$

The log-PL is:

$$l(\beta) = \sum_{t=1}^{N}\{y_t \log \mu_t - \mu_t - \log y_t!\} = \sum_{t=1}^{N}\{y_t Z'_{t-1}\beta - \exp\left(Z'_{t-1}\beta\right) - \log y_t!\},$$

And the equation system of null gradient given below is the first condition of the optimization problem:

$$S_N(\beta) = \nabla l(\beta) = \sum_{t=1}^{N} \{Y_t - \exp(Z_{t-1}'\beta)\} Z_{t-1} = \mathbf{0}.$$

The MPLE $\hat{\beta}$ is the solution of the equation system. The covariance matrix can be checked out in Section 1.4.2.1 of [23].

Remark: Remember that the canonical link is written as follows $g(\mu_t) = \theta_t(\mu_t) = \log \mu_t = \eta_t = Z_{t-1}'\beta$. Estimating β is thus equivalent to estimating θ.

For a Poisson regression, the dispersion coefficient is assumed to be $\hat{\phi} = 1$, simply because the conditional mean and variance are equal, $\mu_t = \sigma^2$. For the record, in regression cases such as Gamma, if ϕ is unknown, we can easily estimate it by the method of moments, which can be used in numerical applications [24]. The formula generated by this method is stated as follows:

$$\hat{\phi} = \frac{1}{T-p} \sum_{t=1}^{T} \frac{\omega_t (Y_t - \hat{\mu}_t)^2}{V(\hat{\mu}_t)}$$

The final goal of data modeling is to obtain fitted values, $\hat{\mu}$, for the mean of response values y. They are compared to observed data afterwards. Prediction intervals can also be furnished under the asymptotic normality assumption as it was shown in Section 1.4 of [23].

2.2.3. Statistical learning methods and diagnostics

One may encounter a lot of models when one tries to model a specific behavior of some observed data. Nevertheless, not all these models have a good quality of adjustment and/or prediction. There exist some statistical learning tools that we will make use of to compare between models. We do not claim that they are the best, but empirical evidence convinces us in our choice: generated results are said to be stable [27]. In order to be assured of our decision-making, we compare one model to another by at least 2 selection criteria. If there is any incoherence, then a 3rd criterion is in addition employed.

2.2.3.1. Testing Hypotheses

In a model, we might want to test the impact of some covariates in the regression model. For this purpose, we test the hypothesis that the value(s) of the covariate(s) is /are zero. If this hypothesis is accepted, we decide that the covariate does not contribute to the regression and can therefore be omitted. Formally, we write:

$$\begin{cases} H0: \boldsymbol{p}(\beta) = \mathbf{0} \\ H1: \boldsymbol{p}(\beta) \neq \mathbf{0} \end{cases}$$

where $p \in \mathbb{R}^p$ is a vector-valued function of β (Definition 1.55 of [23]). The simplest case is when $p \in \mathbb{R}$, i.e., $H0$: $\beta = 0$ against $H1$: $\beta \neq 0$.

The Wald statistic is commonly used for testing this kind of hypothesis set:

$$\omega_N = N\boldsymbol{p}'(\hat{\beta})\left[P'(\hat{\beta})G^{-1}(\hat{\beta})\boldsymbol{P}(\hat{\beta})\right]^{-1}\boldsymbol{p}(\hat{\beta}).$$

The distribution of this test statistic converges to a chi-square random variable with r degrees of freedom. "r" is the number of estimated parameters when the hypothesis $H0$ is true (Section 1.5 of [23] for more details). To sum up, under certain assumptions and hypothesis $H0$, the test statistic ω_N has the same asymptotic distribution, i.e., chi-square with r degrees of freedom.

2.2.3.2. Diagnostics

"Diagnostics in regression analysis consists of procedures for exploring and testing the adequacy and goodness of fit of fitted models" [23]. The examination of several types of residuals and their error, *deviance analysis* via the *scaled deviance*, and the selection criteria *AIC* and *BIC* define the core of statistical diagnostics.

- **Deviance**

It is considered to be one of the most common measures of goodness of fit. A comparison is made between a *saturated* model and a *reduced* one. The saturated model satisfies the following condition: the dimension of the parameter set equals the dimension of observed data. The reduced model, on the other hand, has a smaller parameter set dimension than the saturated model. Let us denote $l(y; y)$ the maximum log-PL corresponding to the saturated model, and $l(\hat{\mu}; y)$ the maximum log-PL of the reduced model. Given that $l(y; y) \geq l(\hat{\mu}; y)$ for exponential family models, the statistic, called the *scaled deviance*, is constructed as follows (Definition 1.65 of [23]):

$$D \equiv 2\{l(y; y) - l(\hat{\mu}; y)\}.$$

This statistic depends on ϕ, a known and consistently-estimated scale parameter. More or less due to asymptotic approximation arguments (check Section 1.6.1 of [23] for proof), the random variable D follows the chi-square distribution with $N - p$ degrees of freedom, say, $D \sim \chi^2_{N-p}$. The *deviance* is nonetheless but a rescaled statistic of the previous one, given by ϕD, and is independent from ϕ. Generally, the smaller the deviance, the better the model goodness of fit.

Model Selection Criteria

Akaike's information criterion (AIC) has demonstrated its good qualities in measuring goodness of fit in empirical applications. AIC is a function of the number of independent model parameters (Definition 1.67 of [23]):

$$AIC(p) = -2\log PL\left(\hat{\beta}\right) + 2p,$$

where $\hat{\beta}$ is the maximum PL estimator of β and $p = \dim(\beta)$, the *model order*. The model that minimizes the previous formula is then selected. Between two models, we choose the one having the smaller AIC. To consistently estimate the model order in most cases, *Bayesian information criterion* (BIC) is used instead of AIC. We employ Schwarz's formulation (Definition 1.68 of [23]):

$$BIC(p) = -2\log PL\left(\hat{\beta}\right) + p\log N.$$

Similar to AIC, we choose the model that has the smallest BIC.

One would see that the likelihood increases when one adds parameters, but this could lead to an overfitting. Both BIC and AIC try to resolve this problem by introducing a penalty term related to the number of parameters in the fitted model; the penalty term is larger in BIC than in AIC. Thus one would use BIC, when a decision based on AIC remains doubtful.

Residuals

A *residual* stands for a certain deviation of a fitted value from an observed value. By analyzing residuals, one can evaluate the goodness of fit of some model compared to the observed data, and may know the significance of covariates on the response variable. Different types of residuals have been proposed in the statistical literature. Deviance residuals, among others, will be used in the context of "Time Series following GLMs", and are formulated as follows (Definition 1.72 of [23]):

$$\hat{d}_t = sign(Y_t - \hat{\mu}_t)\sqrt{2[l_t(Y_t) - l_t(\hat{\mu}_t)]}, t = 1, \dots, N,$$

where the sum of squares of deviance residuals is equal to the scaled deviance statistic.

Error term of a model

In order to compare between models, a prediction error is calculated. For this purpose, the basic residuals, say, observed minus fitted, are calculated for both models. They represent the difference between observed and fitted values and will help us judge over the quality of prediction. In order to calculate this error term, we will simply use the standard deviation for a certain number of fitted values, say, N:

$$Error\ term = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i - \hat{e}_i)^2}$$

The smaller a model's error term is, the better its quality of prediction will be.

2.3. Selecting the model

Regarding the count data, some early conclusions were highlighted and a theoretical introduction was given in the previous sections. We already have knowledge of the two main features of tornado counts: seasonality and upward trend. By using the "Time Series following GLMs" framework, we will include such features as explanatory variables and estimate them using the corresponding techniques of optimization.

Several models will be tested and the selection criteria will be applied. For each one, if it is necessary, a graph will be furnished. This graphical comparison will be necessary to determine whether the model fits the count data well. Some other notions will be discussed too, as they can capture variability above average found in the data. Finally, one of the frequency models will be chosen and used for further study. Let us now list them below.

2.3.1. Classical Poisson (no trend/seasonality)

We start with the homogeneous Poisson distribution, its classical representation. This process is called homogeneous, since the intensity parameter is not adjusted with the time. No trend or seasonality are considered here. In average, we have the same number of events, whenever they occur. In this first model of Poisson distribution, the assumption relying behind is "mean equals variance", so no variability above or below average is considered. In other words, no over-/under- dispersion is taken into account. We write it down as follows:

$$\mu_t = \sigma^2$$

The "Time Series following GLMs" framework, as mentioned above, is used here for the calculations. In this model, there are no covariates in the linear predictor function, only the constant β_0 is kept:

$$\mu_t = \exp(\beta_0)$$

After applying the optimization technique on R, we have the following estimations:

$$\hat{\beta}_0 = 3.23971 \text{ with an error } \widehat{\sigma_{\hat{\beta}_0}} = 0.0093$$
$$\hat{\mu}_t = 50.71$$

In terms of frequency, it means that approximately 50.71 tornadoes occur per month during the 1990-2008 period. In the theory of log-likelihood estimation, for a Poisson process, the estimator equals the empirical mean. We confirm it by calculating the empirical average of the count data, which also is 50.71 tornadoes per month. No dispersion is considered, so the parameter of dispersion is fixed at 1, say, $\hat{\phi} = 1$. The $log - likelihood = -5324.111$, $AIC = 10650$, and $BIC = 10653.65$.

There is obviously a bad fitting with a homogeneous Poisson. Having the same average for all the time period considered is not true in most of the time, such as in the case of the tornado count data. One way to improve this model is by considering the dispersion - variability - found in the data. We should indeed use Poisson processes that are inhomogeneous. More specifically, when a Poisson process assumes an intensity parameter that varies with time, it is said to be inhomogeneous. To strengthen the validity of our point of view, please refer to the detailed statistical description in the beginning of this chapter, which confirmed that the count data have variability 2.5 times above average than below average (also see section 2.3.7). This is thus a strong argument which convinces us of an over-dispersion of the data. In the following models, inhomogeneous Poisson processes will be provided and tested afterwards. Trend, seasonality, combined or taken separately, will be considered for each model. Also, in Chapter 4, the Poisson mean parameter considered in the Poison-multifractal model is time-dependent.

2.3.2. Quasi-Poisson (no trend/seasonality)

Let us first prove the over-dispersion that exists in the arrival data. The same mean representation is proposed as for the Classical Poisson (no trend/seasonality), only the dispersion parameter needs to be adjusted. No trend or seasonality are considered here.

Remember that, with a Poisson distribution, not much can be done, as ϕ is taken for an exogenous parameter. It must so be rendered endogenous, or, put differently, estimable by the model. The family of Poisson distributions appears indeed to be unhelpful in this case. In the statistical theory, another family deriving from the Poisson family has a dispersion parameter, to which is not given a fixed value of 1, and is considered as unknown. We are speaking about the Quasi-Poisson family. The assumption of a unit dispersion coefficient, say, $\hat{\phi} = 1$, is thus not held anymore. We will not go through the details of optimization methods for these quasi-distributions, only a few facts are given below about how they are similar to what was mentioned in the theoretical overview in this chapter.

If one accounts for some sort of trend in the arrival data, one way of including it is by studying a possible mathematical relationship between the conditional mean and the conditional variance [24]:

$$Var[Y_t|\sigma(Z_{t-1})] = \phi * \mathbb{E}[Y_t|\sigma(Z_{t-1})]$$

Remark: In order to avoid any probable confusion, one ought to know that the trend of this relationship must not be interchanged with the time trend component mentioned in Chapter 1 that we will study later.

The GLM framework remains similar as in the case of the "Classical Poisson (no trend/seasonality)" model, except for the dispersion coefficient. The constant β_0 is estimated at $\hat{\beta}_0 = 3.92613$ with an error $\hat{\sigma}_{\hat{\beta}_0} = 0.06526$. In the optimization method, the equation above will be part of the equation set to be solved and an estimate of ϕ belongs to the outputs' summary this time.

The dispersion parameter has a value of approximately 49.24. One may ask how one can interpret its value. For the record, when $\hat{\phi} = 1$, the count data are considered to have no over-/under- dispersion. Here, $\hat{\phi} = 49.24$ shows exactly what we were expecting: high variability exists in the data, and more precisely variability above average. Also, the empirical study of monthly variances in section 2.1.3 showed a strong heteroscedasticity pattern in the data, say, a variance which is not the same from one month to another.

To sum up, the results from the Quasi-Poisson estimation, strengthened by the empirical conclusion of monthly variances, is a mere proof that there is no way a constant intensity would be enough to model such highly volatile data.

For this reason, we will consider other modeling approaches which involve time trend and/or seasonality. These two features can be part of tested models via covariates. For the following nested models under a GLM framework, we consider thus monthly and trend covariates and observe if the covariates are significant and how the log-likelihood is improved from one model to another.

The procedure will be the following:

Each of the following models takes the initial case of a Poisson distribution with constant intensity under a GLM form and add covariates to it.

For each covariate added to the model, we have to check its significance.

A Wald statistic test is performed and the p-value is compared to the alpha risk level, where $\alpha = 5\%, 10\%$ in general. The decision depends on the position of the p-value in comparison with alpha.

For a p-value $< \alpha$, we decide to reject the null hypothesis H0: $\beta_i = 0$ with a 1st error type α, and keep the covariate in the model considered.

Usually, a significant increase in log-likelihood's value demonstrates the importance of adding this new covariate to the model. The higher the log-likelihood is, the better the model will be.

2.3.3. Poisson-trend

This model adds a trend component to the "Poisson (no trend/seasonality)" model. In the linear predictor function, say η_t, we have two coefficients now, β_0 for the constant and β_t for the time covariate. The link function relates the linear predictor to the conditional expectation:

$$\mu_t = \exp(\eta_t) = \exp(\beta_0 + \beta\_t * t).$$

The estimated parameters from the R optimization technique are:

$$\hat{\beta}_0 = 3.49021 \text{ with an error } \widehat{\sigma_{\hat{\beta}_0}} = 0.02083$$
$$\widehat{\beta\_t} = \ 0.00357 \text{ with an error } \widehat{\sigma_{\beta\_t}} = 0.00014$$

In terms of frequency, for example, for time $t = 100$ that corresponds to April 1998, we have an estimated value of approximately 45 tornadoes. The observed number of tornadoes that month was 103.

Adding the time covariate to the model is very significant, because of a p-value = '<2e-16' $\ll \alpha = \ 5\%, 10\%$. The log-likelihood belonging to the "Poisson-trend" model equals -5010.527, which shows an improvement compared to the "Classical Poisson" model log-likelihood: -5324.111.

For information, $AIC \ = \ 10025$ and $BIC \ = \ 10026.48$.

Figure 11 shows us fitted values with a trend component. Apparently, it is insufficiently fitting the data, because it still doesn't show any variability except the upward trend: outliers are not taken into consideration at all.

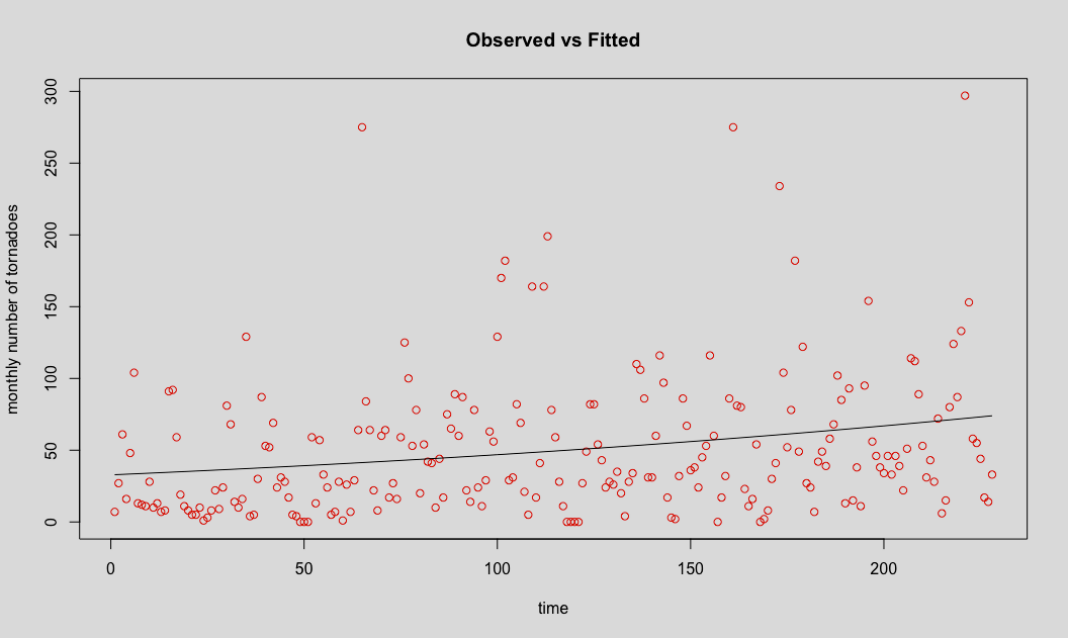

Fig. 11 Fitted values of Poisson-trend model vs observed count data in red

2.3.4. Poisson-seasonality

This time, we consider monthly covariates only, say, an annual seasonality variable with 12 modalities. While working with GLMs, and when we have a variable with more than one modality, it is enough to estimate only $p-1$ modalities - parameters -, where we denote by p the total number of modalities. The remaining modality is called the reference modality and its coefficient is valued to zero. In our case, we consider January to be the reference here, $\beta_1 = 0$:

$$\mu_t = \exp(\beta_0 + 0 * month_1 + \beta_2 * month_2 + \cdots + \beta_{11} * month_{11} + \beta_{12} * month_{12})$$
$$= \exp(\beta_0) * \exp(month_1) * \exp(\beta_2 * month_2) * \ldots * \exp(\beta_{11} * month_{11}) * \exp(\beta_{12} * month_{12}).$$

The modalities $month_i, for\ i = 1, \ldots, 12$, are indicator functions constructed as follows:

$$month_i = \begin{cases} 1\ if\ observation\_mod12 = i \\ 0\ otherwise \end{cases}.$$

μ_t is also expressed as a multiplication of the so-called rating variables in the equation above, which usually are discriminating variables in a non-life insurance framework. In this first model, the modality February only was insignificant with respect to a Wald test, i.e., its p-value = $0.582 > \alpha = 5\%, 10\%$. We reduced in this case this model to a model with seasonality without $month_2$. The coefficient of the modality February is set to $\beta_2 = 0$ in this situation. For this reduced model, the estimated values of the covariate variables and their exponentials are shown in Table 8 below.

Parameters	**Estimation** $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}, \boldsymbol{i} = \mathbf{0}, \ldots \mathbf{12}$	**Standard error** $\widehat{\boldsymbol{\sigma}_{\widehat{\beta}_i}}, , \boldsymbol{i} = \mathbf{0}, \ldots \mathbf{12}$	**p-value**	**exp** $(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}} + \widehat{\boldsymbol{\beta}}_{\mathbf{0}})$ **Monthly estimated**
$\boldsymbol{\beta_0}$	3.22203	0.03239	< 2e-16	-
$\boldsymbol{\beta_1}$	0	-	-	25.08
$\boldsymbol{\beta_2}$	0	-	-	25.08
$\boldsymbol{\beta_3}$	0.77665	0.04488	< 2e-16	54.53
$\boldsymbol{\beta_4}$	1.22925	0.04078	< 2e-16	85.74
$\boldsymbol{\beta_5}$	1.56283	0.03859	< 2e-16	119.68
$\boldsymbol{\beta_6}$	1.11835	0.04165	< 2e-16	76.74
$\boldsymbol{\beta_7}$	0.64147	0.04641	< 2e-16	47.63
$\boldsymbol{\beta_8}$	0.27288	0.05145	1.13e-07	32.95
$\boldsymbol{\beta_9}$	0.46290	0.04868	< 2e-16	39.84
$\boldsymbol{\beta_{10}}$	0.40861	0.04944	< 2e-16	37.74
$\boldsymbol{\beta_{11}}$	0.65246	0.04628	< 2e-16	48.16
$\boldsymbol{\beta_{12}}$	-0.48971	0.06689	2.45e-13	15.37

TABLE. 8 ESTIMATED PARAMETERS AND MONTHLY ESTIMATES FROM POISSON-SEASONALITY MODEL

All the month covariates are now significant in the reduced seasonal model, with their p-values $\ll \alpha = 5\%, 10\%$. As expected, the values of monthly estimations except February, say, the multiplicative rating variables, correspond to the empirical monthly averages given in Table 6. For example, in terms of frequency, we estimate approximately that 120 tornadoes will occur in May.

The $log - likelihood = -3624.734$. In terms of AIC/BIC, with $AIC = 7271.5$ and $BIC = 7314.62$ here, the "Poisson-seasonality" model is better than the "Poisson-Trend" and "Classical Poisson" models.

Figure 12 shows us fitted values of the present model with seasonal components only. Outliers are hardly attained. No trend component is included in this model. We should aim another model accounting for both features.

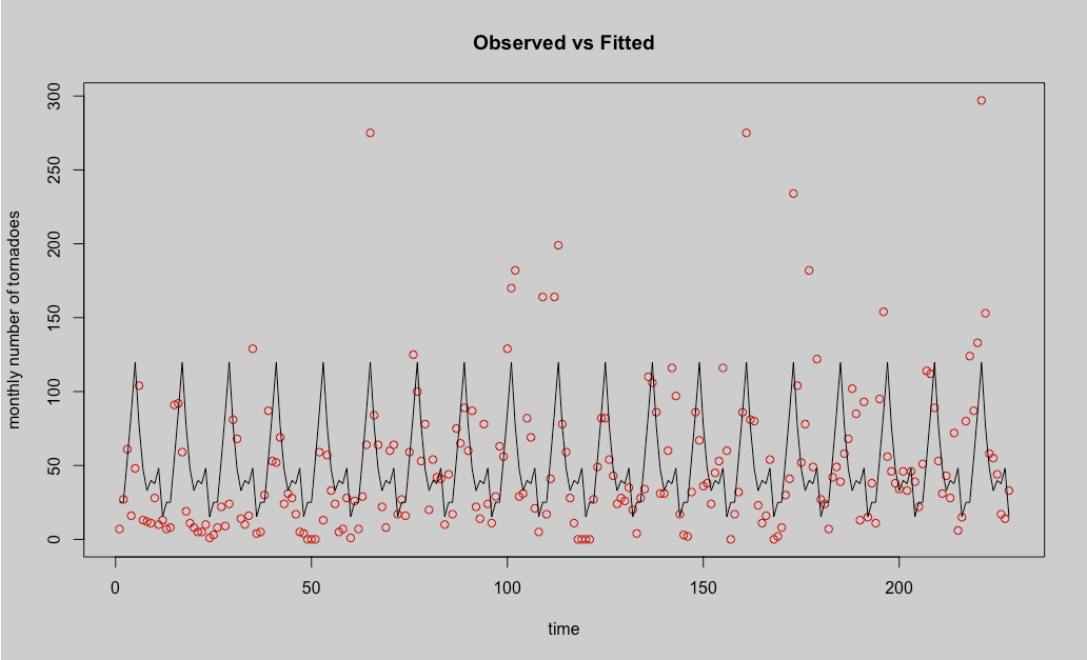

Fig. 12 Fitted values of Poisson-seasonality model vs observed count data in red

2.3.5. Poisson-trend-seasonality

This time, trend and seasonality are considered. This model is a mix between the last three models. The conditional expectation becomes then:

$$\mu_t = \exp(\beta_0 + \beta\_t * t + 0 * month_1 + 0 * month_2 + \beta_2 * month_2 + \cdots + \beta_{11} * month_{11} + \beta_{12} * month_{12}).$$

In this first model, it should be highlighted that the modality February only was insignificant here, with respect to a Wald test, i.e., p-value = 0.543495 > $\alpha = 5\%, 10\%$. The coefficient of the modality February is set to $\beta_2 = 0$ in this situation. We decide to reduce this model to a model with trend and seasonality without $month_2$.

The estimated values of the covariate parameters for the reduced model are shown in Table 9 below.

Parameters	**Estimation** $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}, \boldsymbol{i} = \mathbf{0}, \ldots \mathbf{12}\ \boldsymbol{ou}\ \boldsymbol{t}$	**Standard error** $\widehat{\boldsymbol{\sigma}_{\widehat{\beta}_{\boldsymbol{i}}}}, \boldsymbol{i} = \mathbf{0}, \ldots \mathbf{12}\ \boldsymbol{ou}\ \_\boldsymbol{t}$	**p-value**
$\boldsymbol{\beta_0}$	2.78958	0.03707	< 2e-16
$\boldsymbol{\beta\_t}$	0.00368	0.00014	< 2e-16
$\boldsymbol{\beta_1}$	0	-	-
$\boldsymbol{\beta_2}$	0	-	-
$\boldsymbol{\beta_3}$	0.77113	0.04488	< 2e-16
$\boldsymbol{\beta_4}$	1.22005	0.04078	< 2e-16
$\boldsymbol{\beta_5}$	1.54993	0.03859	< 2e-16
$\boldsymbol{\beta_6}$	1.10178	0.04166	< 2e-16
$\boldsymbol{\beta_7}$	0.62121	0.04642	< 2e-16
$\boldsymbol{\beta_8}$	0.24894	0.05146	1.31e-06
$\boldsymbol{\beta_9}$	0.43527	0.04869	< 2e-16
$\boldsymbol{\beta_{10}}$	0.37730	0.04945	2.35e-14
$\boldsymbol{\beta_{11}}$	0.61747	0.04630	< 2e-16
$\boldsymbol{\beta_{12}}$	-0.52838	0.06691	2.84e-15

TABLE. 9 ESTIMATED PARAMETERS FROM POISSON-TREND-SEASONALITY MODEL

In terms of frequency, for example, for time $t = 100$ that corresponds to April 1998, we have an estimated value of approximately $\hat{\mu}_{100} = \exp(\hat{\beta}_0 + \widehat{\beta\_t} * 100 + \hat{\beta}_4 * 1) = 80$ tornadoes. The observed number of tornadoes during that month was 103.

There is a significant improvement related to the $log - likelihood = -3291.805$. Adding the time covariate known as the trend, it increases the log-likelihood from -3624.734 to -3291.805.

We also have an $AIC = 6607.6$ and $BIC = 6654.191$.

Till now, the "Poisson-trend-seasonality" model is considered the best we have constructed in terms of AIC. Figure 13 represents fitted values of this model with seasonality and trend.

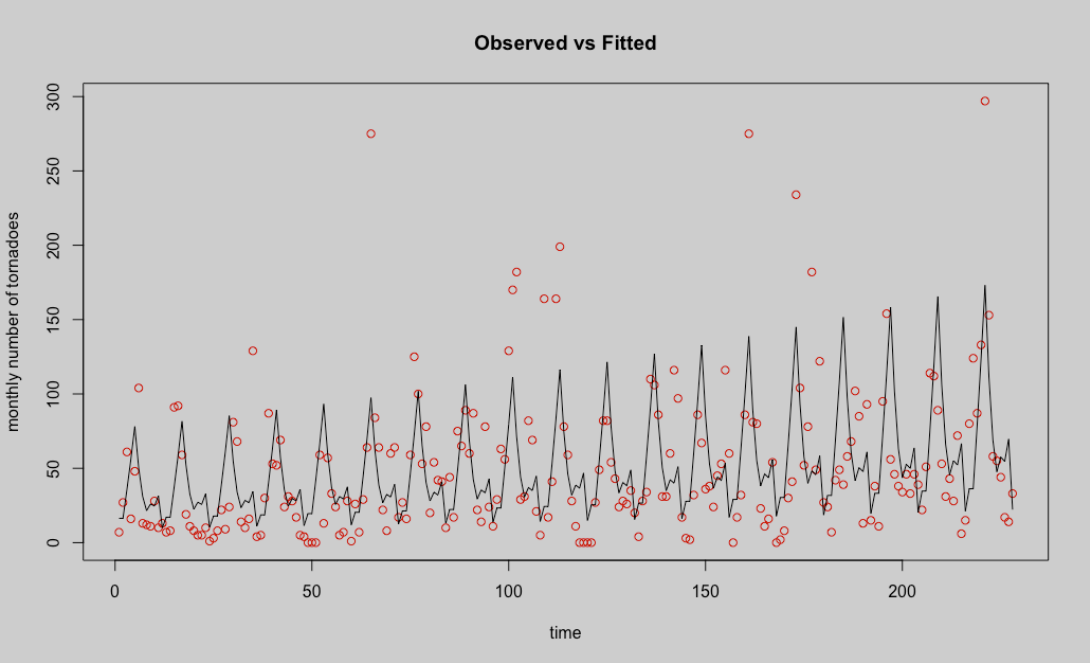

Fig. 13 Fitted values of Poisson-trend-seasonality model vs observed count data in red

Here above, we constructed a model that incorporates time features such as seasonality and trend. One may subsequently ask how one can see if the count data really are dependent on time, say, non-stationary, as we have considered them at last, and if their pattern is representative of this fact. Our purpose here is to mathematically prove, given the observed count data, the empirical conclusions based on trend and seasonality that where highlighted in the early sections of this chapter with the help of STL decomposition of the observed count time series in Section 2.1.2. A simulation set of scenarios will also be used to ensure that the models with such features are robust.

For this purpose, two COSINOR models will be studied below. They imply an alternative way of considering stationary/non-stationary pattern data [25]. It is necessary to specify that these two models won't be considered in the selecting model section, as they serve only for demonstration purposes. Let us now present these 2 methods, opposite to one another. The first method considers data that have a stationary pattern, while the second is about non-stationary pattern data. The definitions of some terms, explicitly taken from the paper, will be developed next to each model below.

2.3.6. Stationary COSINOR

We observed in section 2.1 that the count data have a seasonal pattern. Remember that we also applied "Time Series following GLMs" techniques with a categorical variable of months, which is the parametric model "Poisson-seasonality" in section 2.3.4. This model considers a distribution a priori on the data. Models that assume some parametric seasonal pattern will have a greater power when the parametric model is correct [25]. A popular parametric seasonal model is the COSINOR model [25], which is based on a sinusoidal pattern:

$$Y_t = A * cos(\frac{2\pi t}{c} - P), t = 1\ldots, N,$$

where A is the amplitude of the sinusoidal pattern, or the average height of each seasonal cycle, P is its stationary phase describing the location of the peak of each cycle, c is the length of the seasonal cycle, t is the time of each observation, and n is the total number of times observed.

With a cycle $c = 12$, we model here monthly data with an annual seasonal pattern. The amplitude tells us about the size of seasonal change, and the phase tells us where it peaks [25]. The sinusoid assumes a smooth seasonal pattern that is symmetric about its peak (so the rate of the seasonal increase is equal to the decrease). A and P are the parameters to be estimated. We fit the COSINOR of a Poisson family. We have an estimated parameter $\hat{A} = 0.5151$. As expected, the phase $\hat{P}$ (peak) is in May. The lowest in Dec.

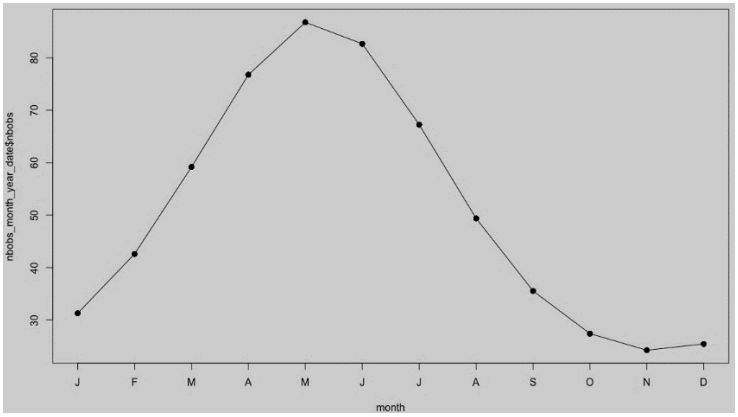

Fig. 14 COSINOR monthly seasonal pattern of the tornado count data

Conclusion: Given the results of the COSINOR model and the graph above, we have demonstrated once again the existence of a seasonal pattern in the count data. Regardless, one may ask now if the data are just stationary, or seasonal, in this case. In Chapter 1, we showed the empirical existence of an upscale time trend. The "Poisson-trend" model

confirmed it too. In the non-stationary COSINOR model, we will not only show the validity of a seasonal pattern, but also the presence of time trend.

Remark: In this model, we have $Log - likelihood = -4219$, $AIC = 8444$. If it were to compare, Models "Poisson-seasonality" and "Poisson-trend-seasonality" would be better than stationary COSINOR, because of a lower AIC.

2.3.7. Non-stationary COSINOR

The seasonal patterns of tornado arrivals gradually change in an upward time trend [25], which means a seasonal pattern that changes from year to year. The model "Stationary COSINOR" illustrated above confirms only partially the validity of this assumption related to stationary seasonality. In such a context, we decide to envisage a non-stationary COSINOR. This model confirms not only the seasonality via the COSINOR structure, but also a present upward trend included in the seasonality. We contemplate thus a non-stationary seasonality feature. The interest shown in this model is thus mainly related to its non-stationary pattern. It reassures us of such a choice as model "Poisson-trend-seasonality".

To fit a non-stationary COSINOR we expand the previous sinusoidal equation of the "Stationary COSINOR" model, thus:

$$Y_t = A_t * cos(\frac{2\pi t}{c} - P_t), t = 1\ldots, N,$$

so that both A_t, the amplitude, and P_t, the phase, are now dependent on time. These parameters represent the non-stationarity feature in the count data.

When a phase is ampler than the previous one, it means that there is an increase of the number of observations in the count data of that phase [25]. The cycle, on the other hand, defines the seasonal pattern of a count time series. In our case, we aim an annual seasonal cycle, so c equals twelve equidistant time intervals.

One may want to know how this model works. The non-stationary COSINOR function uses the Kalman filter to decompose the time series into a trend and seasonal components [25], so it can only be applied to equally spaced time series data, which is our case. The Kalman filter is obtained by using the Markov Chain Monte Carlo (MCMC) algorithm sampling [26]. The frequency estimates are outputted from this algorithm by using the forward and backward sweeps of the Kalman filter, and this is repeated several times. It is indeed very costly in terms of time, as it took us more than half an hour for the R code to show its results. A sufficient number of sample data is needed (we chose 5000) for the estimates to be accurate. We call "burning" the estimates diagnosed as poorly when they are generated from an initial estimation. These last ones are backed out from the code itself. Finally, the function outputs, among others, frequency estimates, a mean trend, a mean season(s), fitted values [trend + season] and 95% confidence intervals per each of

them. Residuals are given too. Other returns are mostly of algorithmic concern on the MCMC method and its residuals.

An R-package has been used for all the calculations. 4000 MCMC samples were needed to reach the best estimates of frequency parameters. A thousand samples sufficed to be considered as "burning". The length of the time series is 228, being the number of observations in the count data. The [trend + season(s)] residuals are quite important to have a thorough look at, since they give us hints on variability above and below average. For this, we extracted this model's output here below.

Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
-70.67	-26.42	-12	-1.448	15.27	184.9

TABLE. 10 NON-STATIONARY COSINOR RESIDUAL STATISTICS OF THE OBSERVED TORNADO COUNT DATA

Although the mean is slightly negative (approximately 1.5 below 0), in absolute values, variability above average is almost 2.5 times more significant than variability below average. In other terms, the data have 2.5 times more chances to increase in numbers than decrease.

Three graphs from the model outputs enforce our statements: the mean trend, the mean season of a 6-month cycle for a comparison, and the mean season of a 12-month cycle. In the second and third graphs below (Figure 15), we can figure out the non-stationary changes in the time trend. We confirm thus the climate change assumptions regarding tornado risk occurrences in the US. The shadows around each drawn line represent the 95% confidence intervals. The third and second graphs of mean season(s) are particularly built around the origin. The pattern does change during the observed period. We expect the changes to be even severer in the years to come.

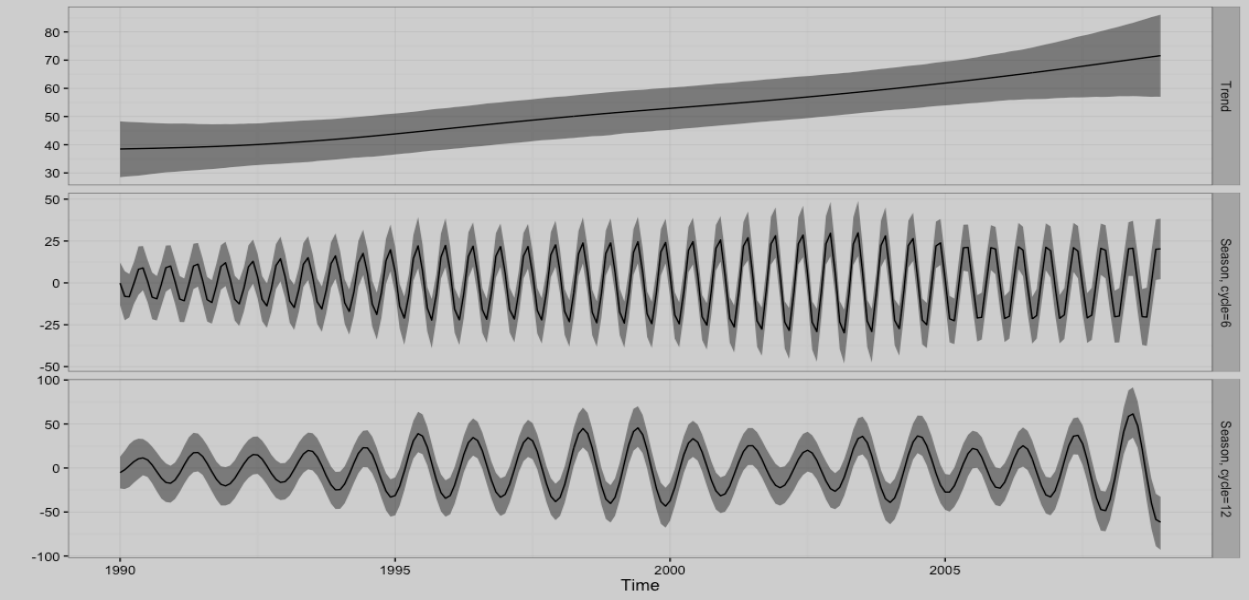

Fig. 15 Non-stationary COSINOR trend and season 6-/12-cycles of the tornado counts

Conclusions:

The validated "Poisson-trend-seasonality" model assumed a non-stationary seasonal pattern. The non-stationary COSINOR model demonstrated the importance of considering the pattern in question as well.

Filters with Markov chains, based on the past information, were also used in the non-stationary COSINOR. We can see that Markov chains per se have a good capacity of incorporating time changes. In parentheses, in Chapter 4, the Multifractal model is developed based on a method that involves them.

2.4. The Frequency model

We have applied models that represent the Classical Poisson (no trend/seasonality), Poisson with a trend pattern, Poisson with a seasonal pattern, and Poisson with trend and seasonal patterns in the previous sections, and a decision needs to be taken for further study in Chapters 4 and 5. On the other side, the "Quasi-Poisson" and "Stationary/Non-Stationary COSINOR" models aren't considered for the selection procedure, as they were only given for study purposes. Remember that the "Quasi-Poisson" model, based on a quasi-likelihood estimation, was only used to show the severe over-dispersion of the count data, say, their empirical variability above average. The two COSINOR generated models helped us confirm the importance of seasonal and trend components in the data.

Regarding the models that do interest us, let us present all the useful information in the following Table 11.1: per each model in a row, we have the number of parameters, the log-likelihood, the model error for a backtesting period 2001-2008, and the criteria needed for the final decision, say, the AIC/BIC.

Section	Model Poisson-	No. of param.	Log-likelihood	Model error	AIC value	BIC value
2.3.1	(no trend/seasonality)	1	-5324.1	49.9	10650	10653.7
2.3.3	Trend	2	-5010.5	48.5	10025	10026.5
2.3.4	Seasonality	11	-3624.7	40.8	7271.5	7314.6
2.3.5	Trend-seasonality	12	-3291.8	38.7	6607.6	6654.2

TABLE. 11.1 FREQUENCY MODEL COMPARISON

As we specified in the theoretical preliminaries, the best model among these four is the one that minimizes the AIC/BIC value. Apparently, the "Poisson-trend-seasonality" model results markedly the best between them according to the selection criteria. It outperforms all the other models; "Poisson (no trend/seasonality)", "Poisson-trend", and "Poisson-seasonality". This constitutes an accurate model choice as well, because it has the smallest error term, which ensures us that its quality of prediction is better than in the case of the other frequency models. We consider in particular two important features that have been highlighted several times in this thesis till now: the trend and seasonality. By the help of empirical and demonstrative conclusions, these conclusions also were enforced.

Table 11.2 (equivalent to Table 9) below regroups all the estimated parameters of the chosen model given by the GLM optimization technique.

Description	Parameters	Estimation $\widehat{\beta}_i, i = 0, \dots 12$ ou t	Standard error $\widehat{\sigma_{\widehat{\beta}_i}}, , i = 0, \dots 12$ ou _t	p-value
Constant	β_0	2.78958	0.03707	< 2e-16
Trend	β_t	0.00368	0.00014	< 2e-16
Seasonality (Jan)	β_1	0	-	-
Seasonality (Feb)	β_2	0	-	-
Seasonality (Mar)	β_3	0.77113	0.04488	< 2e-16
Seasonality (Apr)	β_4	1.22005	0.04078	< 2e-16
Seasonality (May)	β_5	1.54993	0.03859	< 2e-16
Seasonality (Jun)	β_6	1.10178	0.04166	< 2e-16
Seasonality (Jul)	β_7	0.62121	0.04642	< 2e-16
Seasonality (Aug)	β_8	0.24894	0.05146	1.31e-06
Seasonality (Sep)	β_9	0.43527	0.04869	< 2e-16
Seasonality (Oct)	β_{10}	0.37730	0.04945	2.35e-14
Seasonality (Nov)	β_{11}	0.61747	0.04630	< 2e-16
Seasonality (Dec)	β_{12}	-0.52838	0.06691	2.84e-15

TABLE. 11.2 ESTIMATED PARAMETERS OF FREQUENCY MODELS

Conclusion:

We certainly aim, among others, a good fitting in order to predict future values. However, it must be known that our problem does not only concern the quality of prediction related to future values of the count data, but also the selection of an appropriate model with significant monthly/trend covariates via an inference method. As the name of this master's thesis clearly states it, our goal is to confirm the importance of such components in a Cat model. One chooses indeed a model and its features and the same time.

That is what we tried to attain with this first part of our work. We gauged the frequency element and designed its model with seasonal and trend features found in the observed data, which were confirmed as highly significant by their corresponding Wald tests. The estimated values of monthly tornadoes, generated from the "Poisson-trend-seasonality" frequency model, represent a set of simulated values that will help us evaluate tornado insured risk in Chapter 5 with derivative-reinsurance instruments. For the record, tornado risk's hazard component has also a second element, the damage severity. This is the goal of the next chapter. A similar approach of model selection procedure will be undertaken and a severity model will be chosen in the end.

3. Severity - Continuous distributions

The tornado risk profile has a second component still undeveloped in this thesis. For the record, in the definition of hazard that we stated in Chapter 1, weather extreme occurrence and its damage severity are fundamental when evaluating a risk. It was further shown that two supplementary components related to the severity, namely, exposure and vulnerability features, have usually been found in commercialized Cat models. Their methods mostly use a simulation methodology to predict future stochastic events and their related costs. In Chapters 4 and 5, we will tackle this issue and give more details about it. Per contra, we won't be following an exposure-vulnerability model schema here. This is in fact only one way of modeling Nat Cat costs. Actually, the severity approach that we will follow consists in a fully probabilistic schema.

In Chapter 2, we realized a probabilistic study for the frequency model of the count data. In this chapter, a similar schema will be applied to the claims average cost. We may choose to model the average cost or the aggregate cost of claims itself. Indeed, it is mathematically the same, as one can be expressed in terms of the other. After having carefully studied the properties of a Frequency-Severity model, we found the reason why there is no difference, because both result in the same conclusions. This is valid under some assumptions in our thesis: A demonstration is furnished in the Appendix A.

Leading to a model with a distribution that can fit claims data well and that has a good quality of prediction will be our two main goals. Although the severity approach remains similar to Chapter 2's, the "Count Time Series following GLMs" framework wouldn't be the right one for a severity model. As it may not be obvious, we prefer giving a brief clarification on that. We are now dealing with real numbers data, say, cost observations. Discrete distributions such as Poisson processes cannot implement this behavior in the corresponding model. The number of tornadoes was modeled using discrete distributions. In contrast to the count data, we shall aim continuous distributions for the claim costs in the GLMs. The last ones are applicable to a set of observations that can take values in a continuous range such as real numbers, for instance [27].

3.1. A first overview of the claims data

If we refer to the database specifications back in Chapter 1, we decided to take the average cost of observations for our study. It is the third variable in Table 4 with the observed data from SHELDUS. The average cost is given here per each month during the 19-year period. One would need more specifications here about what these costs represent for an (re)insurer. In section 1.3, we stated that the costs in question represent direct property and crop losses, excluding losses from business interruption. It is appropriate to use these amounts, which are necessary to evaluate insured risks from a (re)insurer point of view.

3.1.1. Statistical description of the average cost data

At this stage, let us visualize the monthly average costs in Figure 16 and have a closer look at their statistical features.

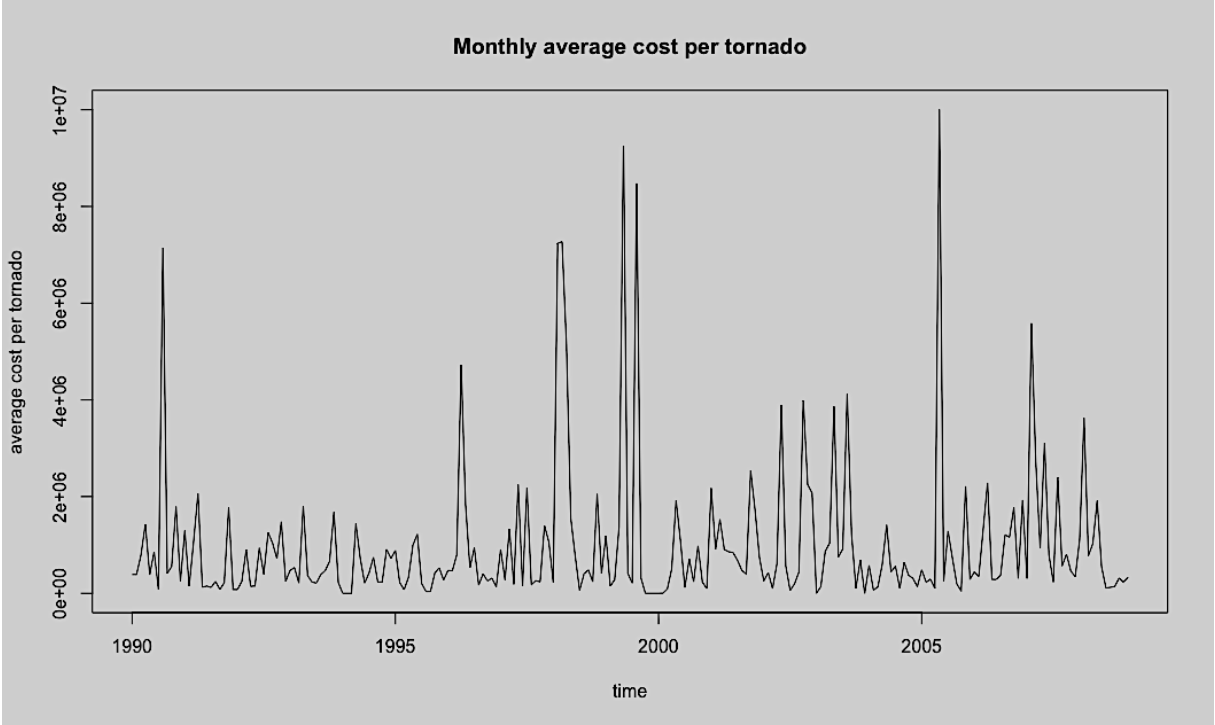

Fig. 16 Monthly average cost of the tornado claim data

Some statistical characteristics of the claim data are made available through the following summarized table. R and SAS software are used to output all the following results. The monthly average cost per tornado in the US, being of $993,800, is given in Table 12. As for the graph, one can observe some extreme average costs per tornado, with the highest peak being of $10,000,000 in 2005. By the use of a statistical learning tool, the boxplot shown below, we can also deduce the severe skewness on the right. In other terms, there are great deviations above the monthly average cost. As a result, we should expect very important variability in the claims data. As it can finally be seen, there are some parallel conclusions between the count data in Chapter 2 and the cost data here. It is the case of the seasonality, for instance.

Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
0	229500	467000	993800	1069000	10000000

TABLE. 12 SUMMARY STATISTICS OF THE TORNADO CLAIM DATA

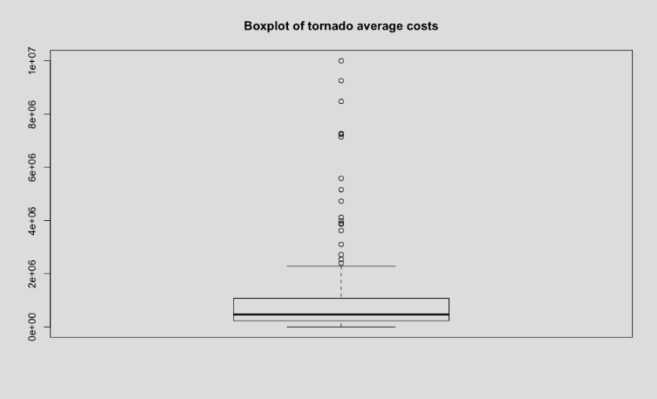

Fig. 17 Boxplot of the tornado claim data

Following per se the same path as we did with the count data, we found it necessary to furnish some preliminary statistics, such as the empirical monthly averages in Table 13. For instance, the monthly average in May is calculated with May average cost observations only. The graph in Figure 18 helps us visualize this statistic. For a clearer view, a graph of monthly variances is shown in Figure 19. We can observe in Table 13 that the peak takes place in February.

Jan	Feb	Mar	Apr	May	Jun
593,254	1,105,150	1,081,055	1,402,821	2,363,261	580,946

Jul	Aug	Sept	Oct	Nov	Dec
483,553	1,530,910	460,099	761,819	1,070,601	492,064

TABLE. 13 AVERAGE COST OF THE TORNADO CLAIM DATA

Also, if we focus on Figures 18 and 19 below and see how the monthly average varies, an annual seasonality may be considered for further modeling. The variability of monthly averages here implicitly incorporates somehow the variability of the observed data too.

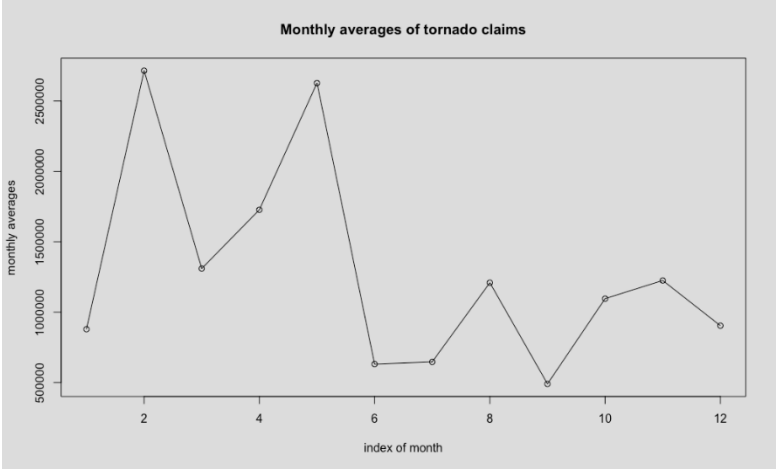

Fig. 18 Monthly average cost of the tornado claim data

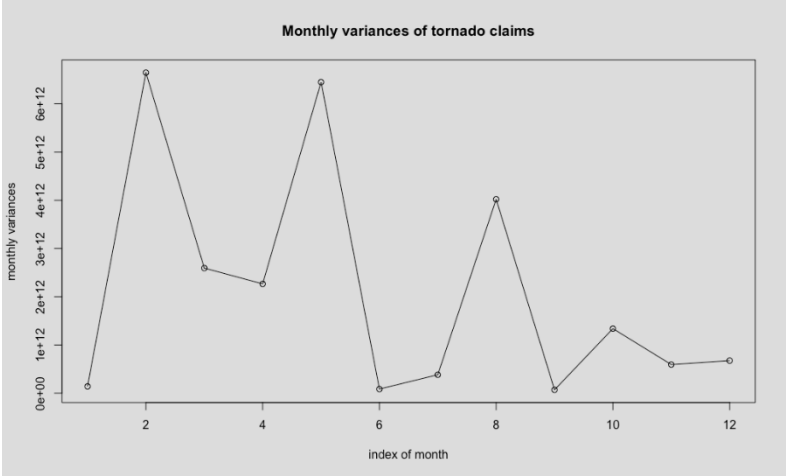

Fig. 19 Monthly variances of cost the tornado claim data

3.1.2. Skewness feature and right-tailed distributions

The skewness plays an important role in our analysis too, and the boxplot confirms it well. The first and second moments above, i.e., mean and variance of a continuous distribution, are not sufficient to fully describe the behavior of the distribution. Better say, they are not capable of capturing finer features as the shape of the probability distribution tail. The tail basically represents values, both negative and positive, that are far away from the mean zone. And so it is in the tail that one can "find" weather extremes' severity [28]. When one refers to the negative values of a distribution far away from the mean zone, they are considered as the left tail. Analogously, the same applies to positive values and the right tail. The cost data is positive and real-valued, and we are as a result interested in a right tail's behavior alone. One might subsequently want to know how a probability tail "behaves". Mathematically speaking, the slower a probability curve decreases, the heavier its tail will be, and so more probable it will be for events causing

such high positive values to happen [28]. One way of directly measuring the heaviness of the tail is by measuring the probability area under the curve. For the same time interval, a heavier-tailed distribution has a higher probability.

The skewness quantity that we mentioned before is one statistically recognized way of measuring the heaviness of the tail. To define it, for a random variable X with mean μ and standard deviation σ, the skewness is the third moment of the centralized and normalized random variable X:

$$\gamma = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

In the case of the average cost observations, the Skewness equals 3.44. In order to interpret this result, there is a conventional rule that states: a distribution is right skewed (left skewed) when its skewness indicator is greater than 3 (lower than -3). Figure 20 visualizes the two cases of skewness in question. As it was clearly seen in Figure 18, there are few average claim amounts with very high values in the observed data, while most of them are low-valued data. The data are indeed right-skewed. It is then in our best interest to choose a distribution with a right-tailed distribution that involves this particular behavior.

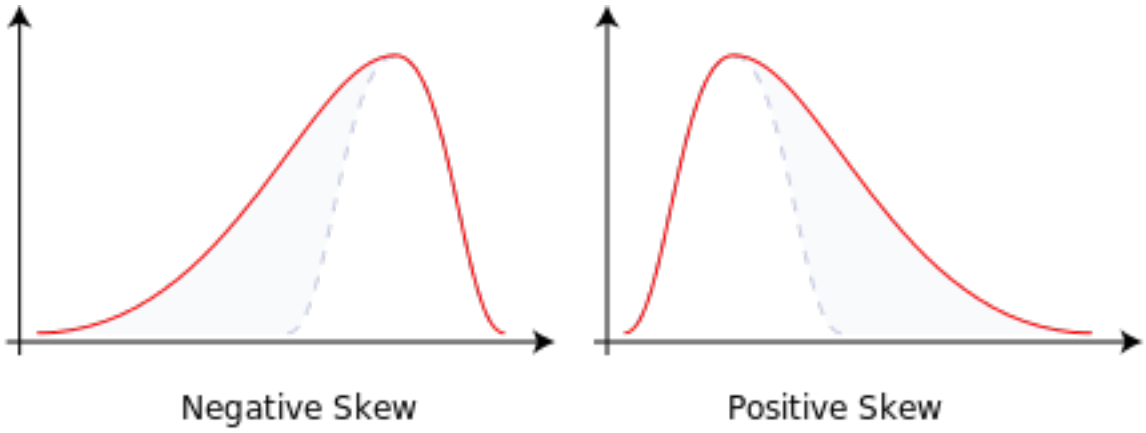

Fig. 20 Skewness [7]

Several scientific papers have proposed models or entire theoretical frameworks to model right-tailed distributions. Some of them also tried to model low-valued and very high-valued data separately. In a (re)insurance context, we would be talking about attritional versus extreme average claims. In spite of that, there will be no separation of claims in this study, since it is not our purpose to reapply the models nor contest the veracity of the aforementioned scientific papers.

[7] Source: Wikipedia

Regarding the work in this Chapter, we will first be interested in finding which right-tailed distribution among those to be tested fits the claims data well. Afterwards, the theoretical framework related to the chosen distribution will be detailed. Thereafter, a list of models similar to those in Chapter 2 will restructure the mean parameter of the chosen continuous distribution, which this turn will improve the log-likelihood. The model error is furnished too. For the sake of further studies in Chapter 5, only one model will be chosen in the end, and it will represent the Severity component of tornado risk.

One ought to know that one is considering here a good model fitting, and significant variables regarding seasonal features as well. This dichotomy makes it more difficult to ensure the efficiency of a model. Some statistical quantities, such as the error term or Wald statistics, will be used to facilitate our reasoning. We will conclude in the end of this chapter about how we will use the chosen severity model later in the following chapters.

3.2. Choosing the right-skewed distribution

We are concerned with finding a theoretical right-skewed distribution that fits the cost observations well. There always is a desire to fit a known probability distribution with reasonably tractable mathematical properties to claims data [29]. A vast palette of known and frequently-used statistical distributions come to our mind. We will focus on the following right-tailed distributions: Exponential, Gamma, Pareto, Weibull and Log-Normal; all somehow easily interpretable and estimable by maximization methods. Our only goal here is to have a fairly good right-skewed distribution and use it later in the severity models. For information, Exponential and Pareto distributions have already been eliminated before in another scientific paper regarding the tornado data [4].

We will be comparing Gamma, Weibull and Log-Normal. The final decision was boosted by these steps in the actuarial modeling process [29]:

- First, a model family is selected from which a distribution in particular is to emerge as the best fitted distribution for claims data.
- Parameters are estimated using the maximum likelihood estimation - MLE - method. After which the log likelihood estimates are computed.
- Then, the goodness of fit is tested using the Q-Q plots. This study used the Q-Q plots to check for the goodness of fit of the chosen distribution to the sampled average claim severity. Other qualitative tests may too be put into practice.
- If the selected model does not fit claims data, another distribution is to be chosen and the process starts again.

In a pure property-casualty insurance severity model, one would compare these fittings and decide which fits the best, from a visualized point of view. We have decided to perform 2 kinds of graphical comparison. Let us now perform the procedure in question and lead to the chosen distribution.

3.2.1. Graphical comparisons

3.2.1.1. Comparison by histograms

The histogram in Figure 21 shows the average cost data. All the following distributions curves are positioned on this histogram. The Normal distribution is plotted as well. The reason why we account for this extra distribution is related to its under-skewness, as its skewness is situated between -3 and 3. This fact strengthens the very choice of a right-skewed distribution, which would account for highly valued average costs in the data. To follow with the first comparison, we have shown the estimated parameters per each distribution here below.

Gamma distribution

$Gamma\ (\alpha, \lambda)$ Its theoretical representation is given in section 3.3. The estimated parameters are generated from a fitted Gamma to the data:

$$\hat{\alpha} = 0.4267238$$
$$\hat{\lambda} = 1.006e-06$$

Normal distribution

The theoretical distribution of this known probability law is given by the following equation:

$$N(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

The parameters that we chose are the empirical mean $\hat{\mu} = 993794.3$ and the empirical variance $\widehat{\sigma^2} = 2.31439e+12$ of the observed data.

Log-Normal distribution

We have the following theoretical probability distribution:

$$\ln N(x; \mu, \sigma) = \frac{1}{x * \sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$$

where $N(x; \mu, \sigma)$ is the Normal distribution. This is derived from the following relation:

$$Y \sim N\ (\mu, \sigma) \Leftrightarrow \exp(Y) \sim \ln N(\mu, \sigma)$$

The parameters of this distribution are defined as follows:

$$E(X) = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$$
$$Var(X) = exp\ (2 * \mu + \sigma\text{^}2) * (exp(\sigma\text{^}2) - 1)$$

For the parameter set, we chose the empirical mean and the empirical variance of the log-normal average cost data, 12.65025 and 7.937418 respectively.

Weibull distribution

And finally, we have the following for the Weibull distribution, for α being the shape and k the scale:

$$Wei\ (x; \alpha, k) = \frac{\alpha}{k}\left(\frac{x}{k}\right)^{\alpha-1} e^{-\left(\frac{x}{k}\right)^{\alpha}}, x > 0$$

The estimated parameters from a fitting to the cost data are $\hat{\alpha} = 0.4267238$ and $\hat{k} = 994{,}035.8$.

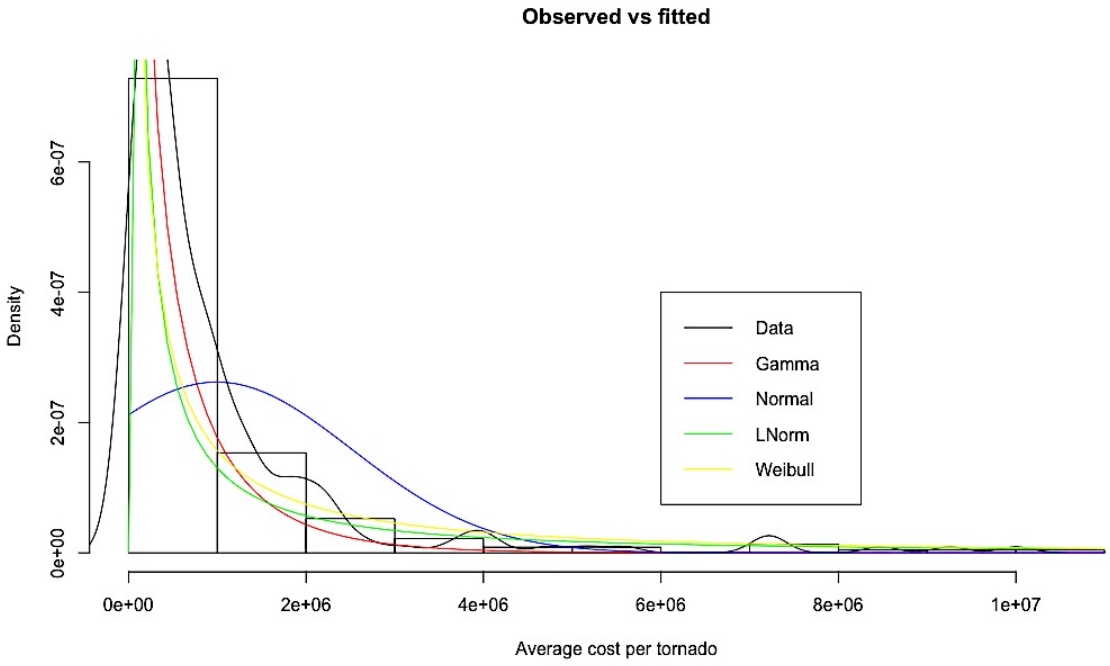

Fig. 21 Histogram comparison of the continuous distributions vs the observed claim data

An early conclusion might be reached: Weibull and/or Log-Normal tend to be the best fitted models in the "heavy-tail" zone, because they include the extreme cost observations. The Normal density is the worst one in fitting the cost data: it is not "enough" right-skewed. To have a closer look at the over/under amplification that a fitted distribution does to the average cost data, let us better refer to the following comparison.

3.2.1.2. Comparison by Quantile-Quantile plots (QQ plots)

There are some aspects of QQ plot in which we are interested here. Their results are quite accurate, as we can also detect outliers, i.e., extreme values, in the graph. In theory, the quantile-quantile (QQ) plot is a graphical technique for determining if two datasets come from populations with a common distribution. A QQ plot is a plot of quantiles of the first dataset against quantiles of the second dataset [30]. In our case, the first dataset would be the tornado average cost observations and the second dataset one of the four fitted distributions. The fact of having a common distribution between the two datasets would be "translated" as follows: the fitted distribution can be used in lieu of the distribution of

the original tornado dataset. Figure 22 plots the original dataset with a 45-degree line. In the following figures 23-26, this will be the reference line and will help us decide whether a fitted dataset is better or worse than another. The greater the departure from this reference line is, the greater the evidence for the conclusion that the two datasets have come from populations with different distributions will be [30]. The graphs appear in the following order: The observed data, Gamma, Normal, Log-Normal, and Weibull.

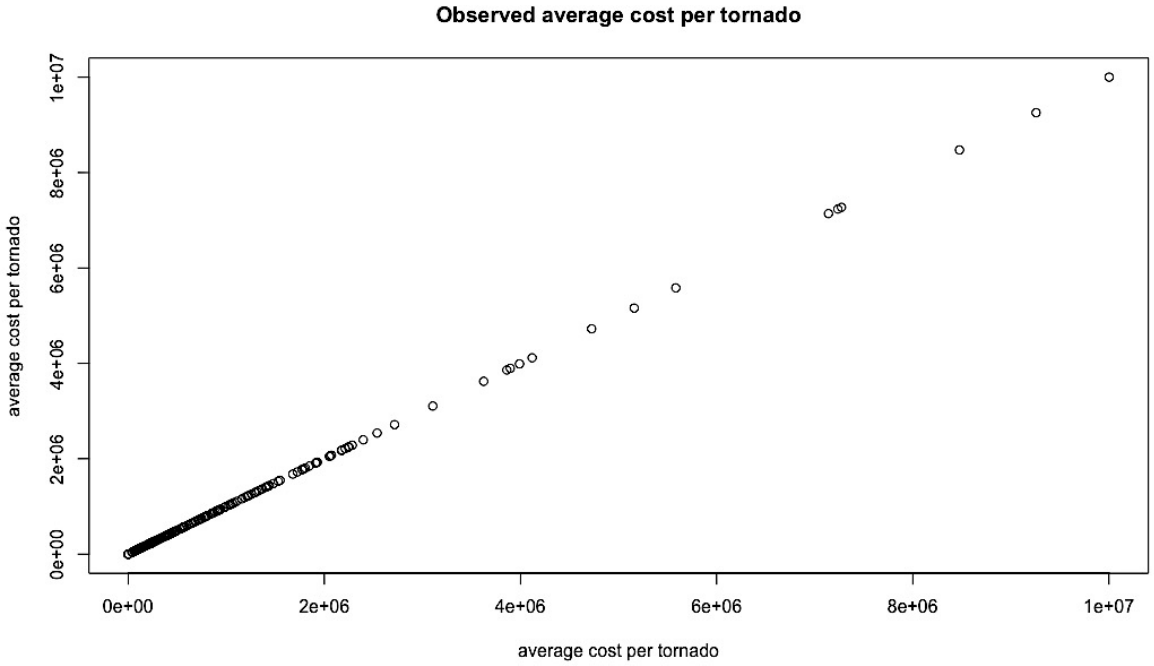

Fig. 22 Observed tornado claim data

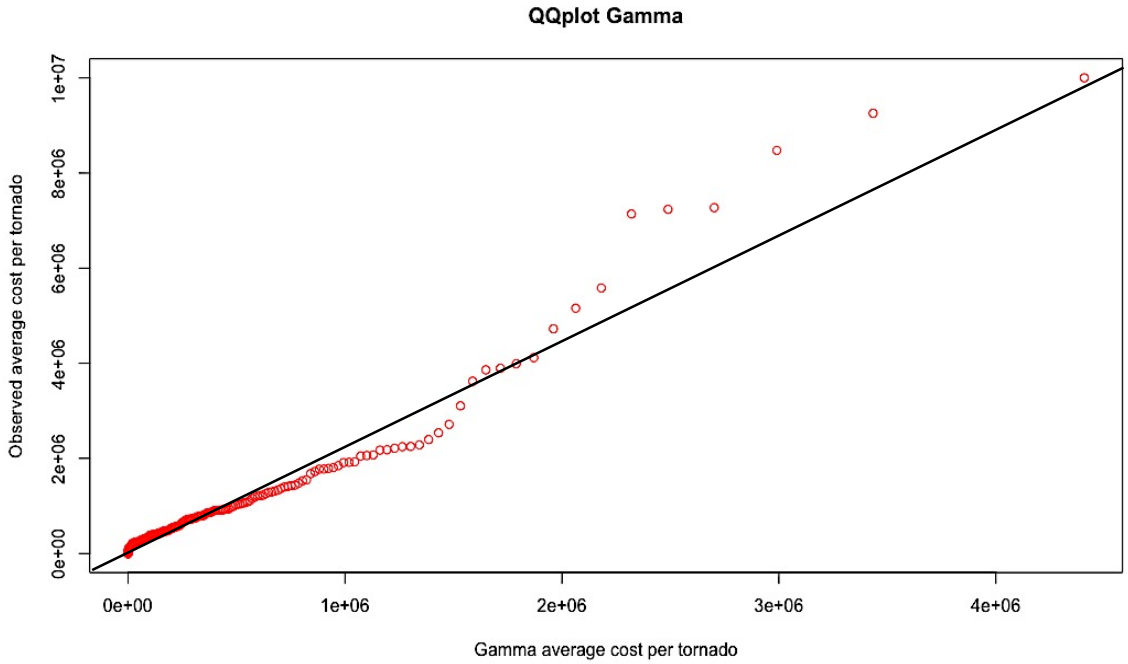

Fig. 23 Fitted Gamma vs the observed tornado claim data

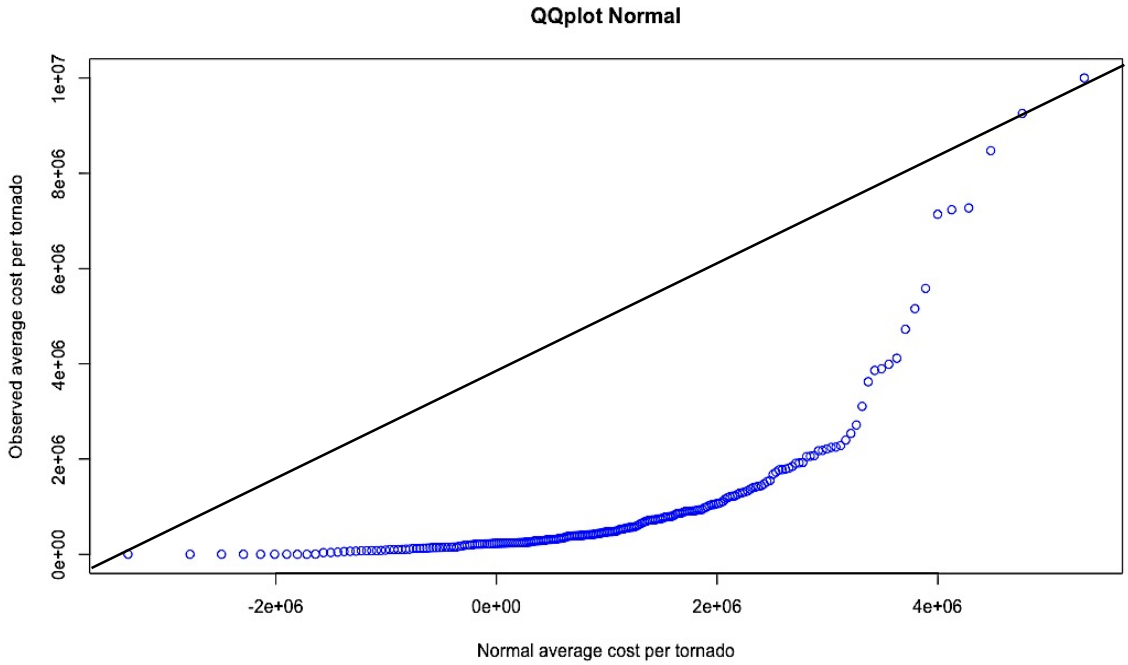

Fig. 24 Fitted Normal vs the observed tornado claim data

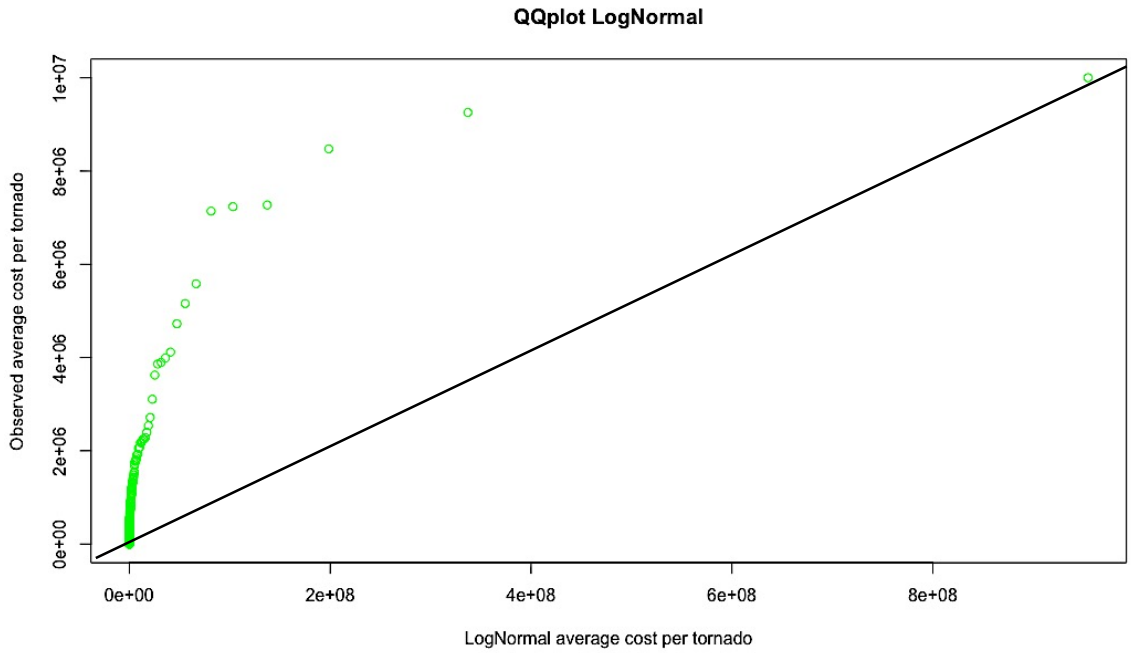

Fig. 25 Fitted Log-Normal vs the observed tornado claim data

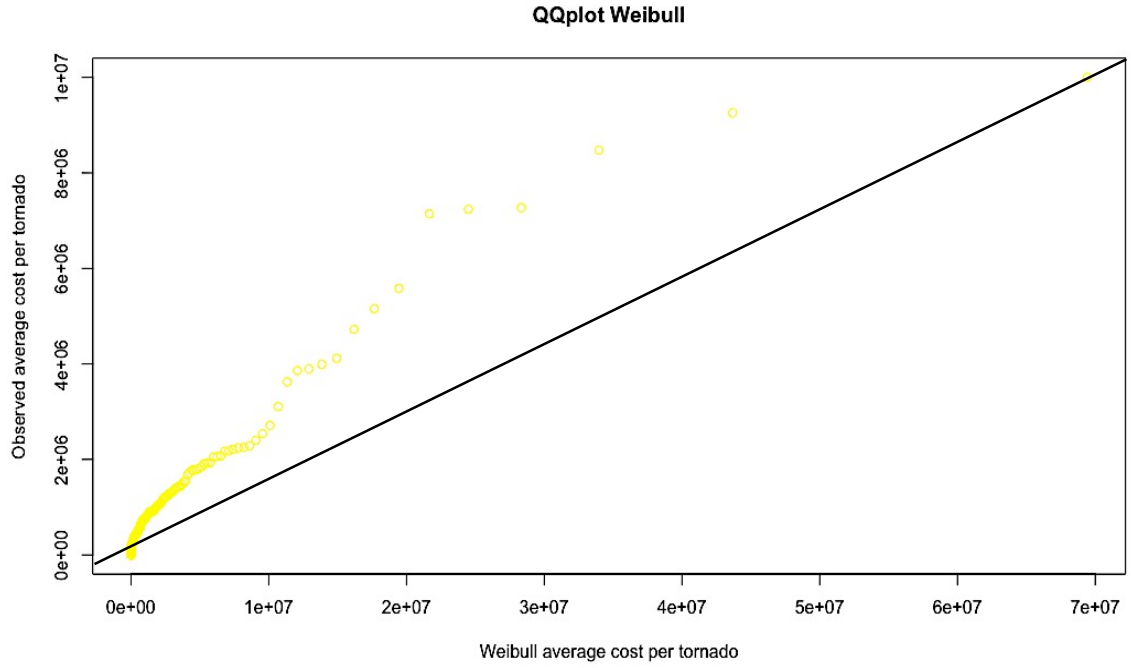

Fig. 26 Fitted Weibull vs the observed tornado claim data

The Gamma dataset seems to be a good fit. Its departure is quite near the 45-degree line, and then sticks to the original dataset till it approaches the 45-50% quantiles. Afterwards, the data become quite rare and we have some very high values, i.e., outliers, which appear to overestimate the original average costs. One could say that there might be a good convergence towards the highly right-skewed zone.

It would be the same in the case of Weibull distribution, as it has a good departure, but quickly overestimates quite a lot the original data starting from the 20% quantile.

On the other hand, Log-Normal and Normal distributions are both bad fitted datasets and appear to be the opposite of each other. The Log-Normal (Normal) has a much greater (lower) departure compared to the 45-degree line and overestimates (underestimates) a lot the original dataset quantiles. Surely, overestimating is more appreciated than underestimating. However, overestimating a lot, as it is the case of Log-Normal, would lead us to very high prices of derivative-reinsurance applications that we put in place in Chapter 5, for instance. None of these last two distributions would be a good candidate for the best fitted distribution.

A qualitative criterion such as a statistical test would be interesting here to choose between Gamma or Weibull. For the record, our interest is based on implementing the GLM framework, as we did in Chapter 2. A distribution has to belong to the Exponential Family in order to apply this framework, discussed in section 2.2.1. In this case, Gamma is the very distribution we have to choose, since it is a member of the Exponential family. On the other hand, the Weibull distribution is not. We decide then to choose the Gamma distribution, member of the Exponential family. Now that we have a continuous distribution for the severity component of tornado risk, we can develop the GLM framework of the Gamma family in the next section before testing the models.

Remark: For information, one would also propose a quasi-distribution, a Generalized Pareto distribution, or other type of distributions, but this would complicate the modeling and the maximization technique by the help of a GLM method, since few right-tailed distributions or none of them are to adapted to such a framework.

3.3. Theoretical preliminaries

We need to focus on inference techniques of interest regarding the comprehension of continuous severity models. Here, we will use the maximum likelihood method, which is one of the most common statistical learning tools to estimate parameters. Let $f(x, \boldsymbol{\theta})$ be the probability density function of a random variable describing a quantitative variable in the population, such as the claims data here. The parameter set $\boldsymbol{\theta}$ is estimated given the sample data $c_1, c_2, \dots, c_N$. The likelihood function given below is used to maximize the likelihood and estimate $\boldsymbol{\theta}$ [23]:

$$L(c_1, c_2, \dots, c_N, \boldsymbol{\theta}) = \prod_{i=1}^{N} f(c_i, \boldsymbol{\theta})$$

Literally, the likelihood of an observed dataset is the probability of obtaining that exact observed dataset given the chosen probability model. Our objective is to find the maximum likelihood estimators - MLE. We calculate the logarithmic function of the likelihood function and maximize it by using numerical methods such as the Fisher Scoring method. Mathematically, we have [23]:

$$\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \sum_{t=1}^{N} \log f(c_i, \boldsymbol{\theta}),$$

MLE has some very good statistical properties, such as uniqueness, existence, consistency and convergence. They were proved in [23].

In the case of Gamma distribution, we have two ways of representing their probability distribution function [31]:

$C^* \sim Gamma\ (\alpha, \lambda)$, where α is the shape parameter and λ the rate[8], has the following probability distribution:

$$f_{C^*}(c) = \frac{\lambda^{\alpha}}{\Gamma(\lambda)} c^{\alpha-1} \exp(-\lambda c) \mathbb{I}_{(0,\infty)}(c)$$

The following transformation is the second way; for $C \sim Gamma\ (\tau, \nu)$, where $\tau = \frac{\alpha}{\lambda}$:

$$f_C(c) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu c}{\tau}\right)^{\nu} \exp\left(-\frac{c\nu}{\tau}\right) \frac{1}{c} \mathbb{I}_{(0,\infty)}(c)$$

and it has a variance of $\frac{\tau^2}{\nu}$. We will especially use the second formulation of Gamma distributions, because it is better when one uses the GLM framework and Chapter 4's techniques. The canonical link function of Gamma distribution is used here by default,

[8] The inverse of the scale parameter φ of Gamma distributions.

and no a priori reason stands actually behind this choice. Anyway, one may find in the statistical literature that the fact of using the canonical link function leads to some desirable statistical properties, such as the simplification of the MLE derivation, the zero sum of residuals, etc. [32].

The second parametrization below allows us to find the following elements of the GLM form for a Gamma distribution:

$$f_C(c) = c(c_t; \phi) exp\left\{\frac{c_t \boldsymbol{\theta} - b(\boldsymbol{\theta})}{\phi}\right\}$$

where $c(c_t; \phi) = \frac{c^{\nu-1}\nu^{\nu}}{\Gamma(\nu)}$, $\boldsymbol{\theta} = -\frac{1}{\tau}$, $\phi = \frac{1}{\nu}$, $b(\boldsymbol{\theta}) = -\ln(-\boldsymbol{\theta})$.

The expected value and variance of the Gamma distribution in the 2nd parametrization case are as follows [32]:

$$\mathbb{E}[C] = b'(\boldsymbol{\theta}) = -\frac{1}{\theta_t} = \tau$$
$$Var[C] = \phi b''(\boldsymbol{\theta}) = \left(\frac{1}{\nu}\right)\left(\frac{1}{\boldsymbol{\theta}^2}\right) = \frac{\tau^2}{\nu}$$

The likelihood function becomes as follows:

$$L(c_1, c_2, \dots, c_N, \boldsymbol{\theta}) = \prod_{i=1}^{N} c(c_i; \phi) exp\left\{\frac{c_i \boldsymbol{\theta} - b(\boldsymbol{\theta})}{\phi}\right\}$$

Finally, the maximization problem is:

$$\max_{\boldsymbol{\theta}} l(c_1, c_2, \dots, c_N, \boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(\ln\big(c(c_i; \phi)\big) + \left(\frac{c_i \boldsymbol{\theta} - b(\boldsymbol{\theta})}{\phi}\right)\right)$$

Remark: Remember that the canonical link is written as follows $g(\tau_t) = \theta_t(\tau_t) = -\ln(-\boldsymbol{\theta}) = \eta_t = Z'_{t-1}\beta$. Estimating β is thus equivalent to estimating θ [23].

Testing hypotheses, goodness of fit estimators, and other diagnostics tools from Chapter 2 remain valid in the case of continuous distributions. We won't become repetitive by developing the theoretical background on these subjects again. One may refer to Chapter 2 if needed. We will apply them to different severity models to decide about their statistical properties. We may now jump to the model selection section.

3.4. Selecting the model

Regarding observed average costs per tornado, some early conclusions were highlighted and the Gamma distribution, which is a good representative of right-skewed distributions and a member of the Exponential Family, was chosen in section 3.2. Additionally, a theoretical introduction on the Gamma family and the application of maximization method were given in the previous section. We already have knowledge of some main features of the tornado claims data, such as the fact of belonging to right-tailed distributions' family whose true distribution is unknown, high variability present in the dataset under the form of outliers, or this upward trend we also captured in the count data in Chapter 2. By using the GLM framework, we will include such features as explanatory variables and estimate them using the corresponding techniques of optimization.

The approach developed here is similar to Chapter 2's model selection. Several models will be tested and the model selection criteria will be applied. For each model, if it is helpful, a graph will be furnished. This graphical comparison will be necessary to determine whether the model fits the data well. Finally, one of the versions will be chosen and used for further study. Let us now list the severity models below.

3.4.1. Classical Gamma (no trend/seasonality)

In this model, we consider the simplest case, maximizing the log-likelihood of a Gamma distribution being fitted to the cost data. For a Gamma distribution with the first representation:

$$Gamma\ (\alpha, \lambda)$$
$$\theta \equiv -\frac{\alpha}{\lambda}$$
$$\varphi \equiv 1/\lambda$$

Where α is the shape, λ the rate, φ the scale, and θ is the estimated parameter. There is only the constant coefficient in the linear predictor. The canonical function is written as follows:

$$-\ln(-\boldsymbol{\theta}) = \eta_t = \beta_0$$

Choosing the Gamma distribution in the step-by-step method exposes us to another problem. This distribution allows positive values only. The presence of zero monthly average claim amounts induced our code to errored results. It is necessary to specify that these zero claim amounts derive from monthly data with no tornado arrivals. We were hence obliged to find a way how to resolve this computing problem. R bloggers proposed at least 3 solutions, of which we chose the very one that eliminates these zero observations from the data, as no tornadoes occurred during these months. This appears to have an insignificant impact on each model's results.

A GLM was fitted and the estimated parameters of this model are:

$$\hat{\alpha} = 0.4267238$$
$$\hat{\lambda} = 1.006e-06 (= \hat{\beta}_0) \text{ with an error } \widehat{\sigma_{\hat{\beta}_0}} = 1.020\text{e}-07$$
$$\hat{\varphi} \equiv \frac{1}{\hat{\lambda}} = 994{,}035.8$$

In the theory of log-likelihood estimation, for a Gamma process, the estimated average given by $\hat{\varphi}$ equals the empirical mean. This is well the case. In terms of severity, the estimated monthly average cost per tornado is $994,035.8 in the 1990-2008 period.

The estimated statistics were given by the optimization technique in R:

$$log - likelihood = -3266.55$$
$$AIC = 6535.1 \text{ and } BIC = 6538.5.$$

One way to improve this model is by considering the variability feature found in the data. Trend and seasonality, all together or taken separately, will be considered for each model.

3.4.2. Gamma-trend

The model adds a trend covariate to the previous one. In the linear predictor function, say, η_t, we have two coefficients now, β_0 for the constant and β_t for the trend covariate. The link function relates the linear predictor to the canonical function:

$$-\ln(-\boldsymbol{\theta}) = \eta_t = \beta_0 + \beta\_t * t$$

The estimated parameters from the R optimization technique were:

$$\hat{\beta}_0 = 1.238e-06 \text{ with an error } \widehat{\sigma_{\hat{\beta}_0}} = 2.280\text{e}-07$$
$$\widehat{\beta\_t} = -1.89e-09 \text{ with an error } \widehat{\sigma_{\widehat{\beta\_t}}} = 1.562\text{e}-09$$

In terms of severity, for time $t = 100$ that corresponds to April 1998, the estimated monthly average cost per tornado is approximately $953,292.9 vs the real average observation cost that is $5,159,128.9.

The constant is significant with respect to a Wald test. However, adding a trend covariate to the Gamma-trend model is not significant, because of a p-value $= 0.227 > \alpha = 5\%, 10\%$. We therefore decide to keep the null hypothesis $H0$ with a 2nd error type β.

The desired outputs from the present model are:

$$AIC = 6535.869 \text{ and } BIC = 6536.454$$
$$log - likelihood = -3265.869$$

There is no such significant difference between "Gamma (no trend/seasonality)" and "Gamma-trend" models. For the record, an important variation of AIC from one model to another is needed in order to rely on such a criterion. A graph with the fitted values is furnished below.

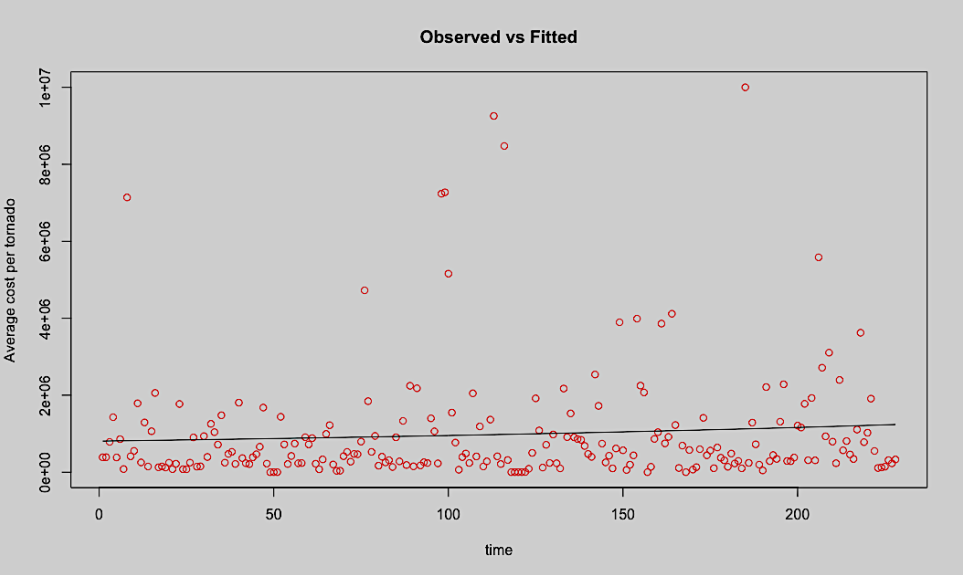

Fig. 27 Fitted values of Gamma-trend model vs observed tornado claim data in red

3.4.3. Gamma-seasonality

This time, we consider monthly covariates only, say, an annual seasonality variable with 12 modalities. While working with GLMs, and when we have a variable with more than one modality, it is enough to estimate only $p-1$ modalities - parameters -, where we denote by p the total number of modalities. The remaining modality is called the reference modality and its coefficient is valued to zero. In our case, we consider January to be the reference here, $\beta_1 = 0$:

$$-\ln(-\boldsymbol{\theta}) = \eta_t = \beta_0 + 0 * month_1 + \beta_2 * month_2 + \cdots + \beta_{11} * month_{11} + \beta_{12} * month_{12}$$

The modalities $month_i$, $for\ i\ = 1, \ldots, 12$, are indicator functions constructed as follows:

$$month_i = \begin{cases} 1\ if\ observation\_mod12 = i \\ 0\ otherwise \end{cases}.$$

An R code is written to model the GLM with monthly modalities. We have the following estimations in Table 14:

Parameters	Estimation $\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}, \boldsymbol{i} = \mathbf{0}, \mathbf{1}, \ldots \mathbf{12}$	Standard error $\widehat{\boldsymbol{\sigma}_{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}}}, \boldsymbol{i} = \mathbf{0}, \mathbf{1}, \ldots \mathbf{12}$	p-value	Monthly estimated in \$ $(-\mathbf{1}/\widehat{\boldsymbol{\theta}})$
$\boldsymbol{\beta_0}$	1.6856e-06	4.655 e-07	0.000367*	-
$\boldsymbol{\beta_1}$	0	-	-	593,254.4
$\boldsymbol{\beta_2}$	-7.807626e-07	5.284 e-07	0.140941	1,105,149.7
$\boldsymbol{\beta_3}$	-7.605947e-07	5.310 e-07	0.153509	1,081,054.6
$\boldsymbol{\beta_4}$	-9.727678e-07	5.055 e-07	0.055605*	1,402,820.5
$\boldsymbol{\beta_5}$	-1.262473e-06	4.800 e-07	0.009147*	2,363,260.7
$\boldsymbol{\beta_6}$	3.571328e-08	6.654 e-07	0.957246	580,945.9
$\boldsymbol{\beta_7}$	3.824090e-07	7.369 e-07	0.604310	483,552.8
$\boldsymbol{\beta_8}$	-1.032411e-06	4.993 e-07	0.039849*	1,530,909.7
$\boldsymbol{\beta_9}$	4.878267e-07	7.597 e-07	0.521440	460,099.3
$\boldsymbol{\beta_{10}}$	-3.729690e-07	5.900 e-07	0.527995	761,818.7
$\boldsymbol{\beta_{11}}$	-7.515626e-07	5.322 e-07	0.159369	1,070,601.0
$\boldsymbol{\beta_{12}}$	3.466383e-07	7.292 e-07	0.635017	492,064.1

TABLE. 14 ESTIMATED PARAMETERS AND MONTHLY ESTIMATES OF GAMMA-SEASONALITY MODEL

In terms of severity, for example, at time $t = 100$ that corresponds to April 1998, the monthly average cost per tornado is estimated to be \$1,402,820.5 during 1990-2008 period vs the real observation that was \$5,159,128.9. It is a better estimation than the "Gamma-trend" model estimation of \$953,292.9.

The calculated estimated statistics are:

$$AIC = 6532.6 \text{ and } BIC = 6573.8$$
$$log - likelihood = -3254.3$$

There certainly is a slight improvement compared to the "Gamma-trend" model. Except for β_0 and modalities $\{month_4, month_5, month_8\}$, all the others are insignificant in respect to a Wald test with an $\alpha = 5\%, 10\%$, and are thus set to 0.

Figure 28 shows us the fitted values of this model with seasonality. Outliers are left outside the predicted values zone. No trend covariate is included in this model. We should aim another model accounting for both features. All the seasonal modalities, except $month_4, month_5, and\ month_8,$ are omitted one by one and we observe the differences in AIC from one model to another.

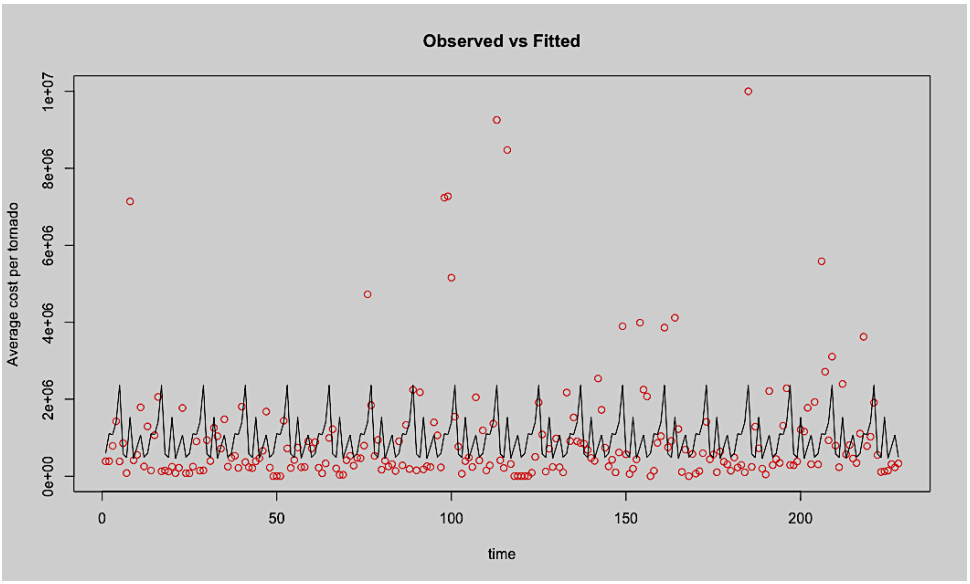

Fig. 28 Fitted values of Gamma-seasonality model vs observed tornado claim data in red

3.4.4. Gamma-seasonality (only months 4, 5, 8)

Following the same approach as in the "Gamma-seasonality" model, we omitted one by one modalities by comparing the nested models in terms of AIC, and finally reached a GLM model only with the β_0 and modalities $month_4, month_5, and\ month_8$. The significance of the other modalities' coefficients was tested. They resulted to be insignificant in section 3.4.5, and so the value of corresponding GLM coefficients are set to 0 here, as it is the case of the reference modality. This model is so far the best in terms of AIC. The linear predictor in this model has the following representation:

$$-\ln(-\boldsymbol{\theta}) = \eta_t = \beta_0 + 0 * month_1 + 0 * month_2 + 0 * month_3 + \beta_4 * month_4 + \beta_5 * month_5 + 0 * month_6 + 0 * month_7 + \beta_8 * month_8 + 0 * month_9 + 0 * month_{10} + 0 * month_{11} + 0 * month_{12}$$

The model delivers the estimated values of parameters:

Parameters	Estimation $\widehat{\beta}_i, i = 0, 1, \dots 12$	Standard error $\widehat{\sigma_{\widehat{\beta}_i}}, , i = 0, 1, \dots 12$	p-value	Monthly estimated in \$ $(-1/\widehat{\theta})$
β_0	1.357765e-06	1.441e-07	< 2e-16	-
$\beta_i, i \in \{1, 2, 3, 6, 7, 9, 10, 11, 12\}$	0	-		736,504.5
β_4	-6.449154e-07	2.689e-07	0.01729	1,402,820.5
β_5	-9.346208e-07	1.973e-07	3.86e-06	2,363,260.7
β_8	-7.045586e-07	2.531e-07	0.00583	1,530,909.7

TABLE. 15 ESTIMATED PARAMETERS AND MONTHLY ESTIMATES OF GAMMA-SEASONALITY (ONLY $month_4, month_5, and\ month_8$) MODEL

The calculated estimated statistics are:

$$AIC = 6525.2 \text{ and } BIC = 6538.92$$
$$log - likelihood = -3258.6$$

Figure 29 show us the fitted values of this model. Outliers are left outside the predicted values zone.

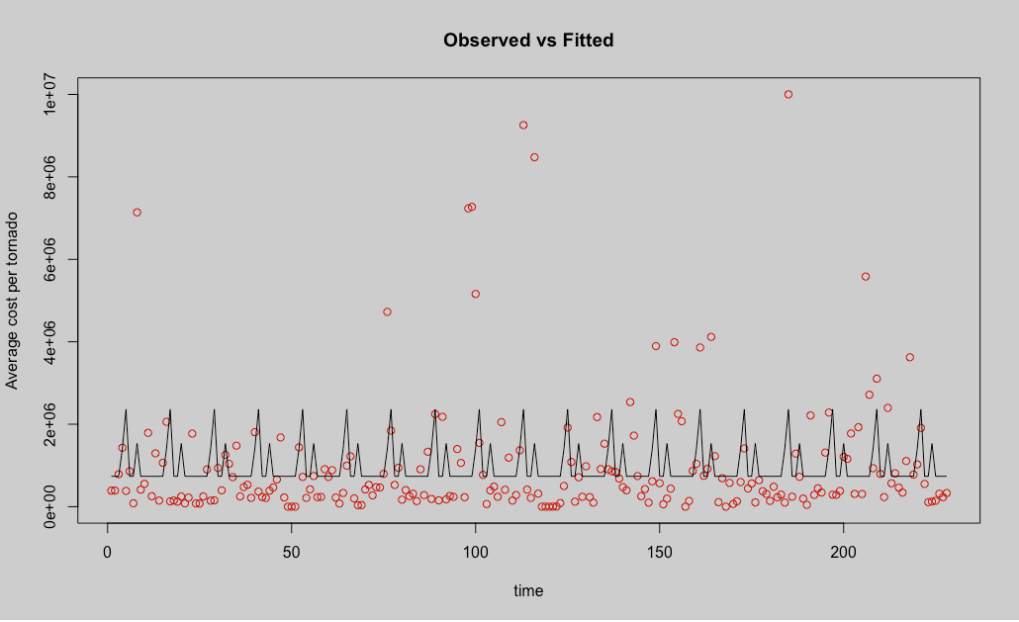

Fig. 29 Fitted values of Gamma-seasonality (only $month_4, month_5, and\ month_8$) model vs observed tornado claim data in red

In respect to a Wald test with an $\alpha = 5\%, 10\%$, all the 3 covariates are significant. This model also is better than the "Gamma-seasonality" model when comparing their AIC and BIC value. Let us implement a last model involving all the features and decide which one to choose for further study.

3.4.5. Gamma-trend-seasonality

This time, trend and seasonality are considered altogether. The trend covariate from the "Gamma-trend" model is added to the "Gamma-seasonality" model. The present model is a mix between the last two models in question. The linear predictor becomes then:

$$-\ln(-\boldsymbol{\theta}) = \eta_t = \beta_0 + \beta\_t * t + 0 * month_1 + \beta_2 * month_2 + \cdots + \beta_{11} * month_{11} + \beta_{12} * month_{12}$$

The coefficient estimates are given below:

Parameters	Estimation $\widehat{\boldsymbol{\beta}}_i, \boldsymbol{i = 0, 1, \ldots 12\ ou\ t}$	Standard error $\widehat{\boldsymbol{\sigma}_{\widehat{\beta}_i}}, , \boldsymbol{i = 0, 1, \ldots 12}$	p-value
$\boldsymbol{\beta_0}$	1.860428e-06	4.853e07	0.000166*
$\boldsymbol{\beta\_t}$	-1.547500e-09	1.089e09	0.156935
$\boldsymbol{\beta_1}$	0	-	-
$\boldsymbol{\beta_2}$	-7.739445e-07	5.310e07	0.146411
$\boldsymbol{\beta_3}$	-7.524765e-07	5.337e07	0.160002
$\boldsymbol{\beta_4}$	-9.598062e-07	5.080e07	0.060182*
$\boldsymbol{\beta_5}$	-1.238255e-06	4.829e07	0.011028*
$\boldsymbol{\beta_6}$	4.332372e-08	6.693e07	0.948450
$\boldsymbol{\beta_7}$	3.905612e-07	7.414e07	0.598903
$\boldsymbol{\beta_8}$	-1.011952e-06	5.020e07	0.045046*
$\boldsymbol{\beta_9}$	4.988316e-07	7.645e07	0.514753
$\boldsymbol{\beta_{10}}$	-3.573028e-07	5.934e07	0.547694
$\boldsymbol{\beta_{11}}$	-7.311717e-07	5.351e07	0.173256
$\boldsymbol{\beta_{12}}$	3.626159e-07	7.338e07	0.621696

TABLE. 16 ESTIMATED PARAMETERS OF GAMMA-TREND-SEASONALITY MODEL

In terms of severity, for example, at time $t = 100$ that corresponds to April 1998, the monthly average cost per tornado is estimated to be \$1,340,710 vs the real observed average that was \$5,159,128.9. It seems to be a more moderated estimation than the "Gamma-seasonality" model with \$1,402,820.5.

The following statistical estimates have been calculated:

$$AIC = 6533.4 \text{ and } BIC = 6577.98$$

$$Log - likelihood = -3253.7$$

With respect to a Wald test an $\alpha = 5\%, 10\%$, adding a time trend covariate does not embed extra significance to the model, and we decide thus to accept the null hypothesis with a 2nd error term β. Except for β_0 and modalities $\{month_4, month_5, month_8\}$, all the others are insignificant in respect to a Wald test with an $\alpha = 5\%, 10\%$. The graph in Figure 30 shown above helps in visualizing seasonality and trend features of the cost data.

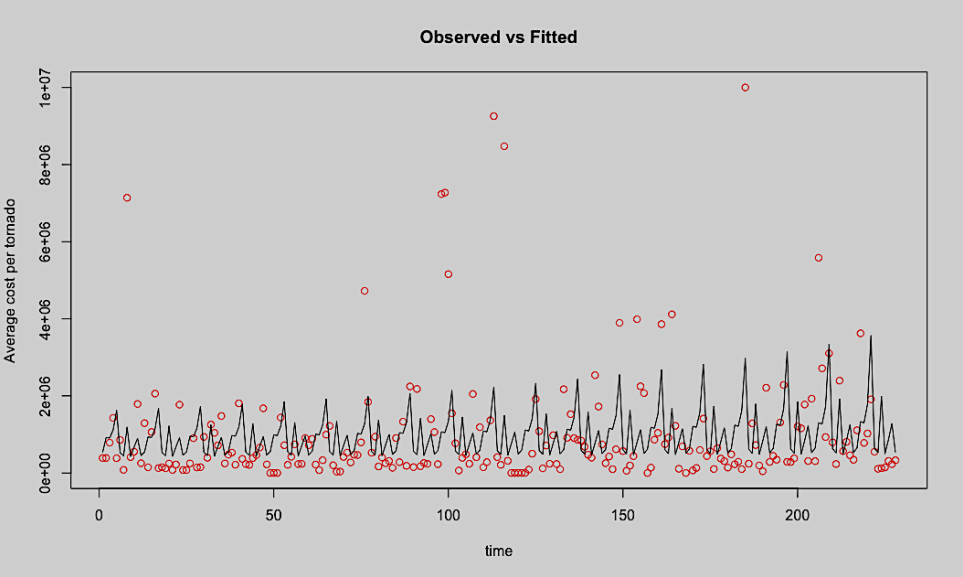

Fig. 30 Fitted values of Gamma-trend-seasonality model vs the observed tornado claim data in red

3.5. The Severity model

To summarize the previous sections, we have applied the following models: the classical version of a Gamma distribution (no trend/seasonality), Gamma with a trend pattern, Gamma with a seasonal pattern, and Gamma with trend and seasonal patterns. A decision needs to be taken for further study in Chapters 4 and 5. Regarding each of one these models, let us present the corresponding statistical information in Table 17: number of parameters, error term, and model decision criteria.

Section	Model: Gamma-	No. parameters	Log-likelihood	Model error	AIC value	BIC value
3.4.1	(no trend/seasonality)	1	-3266.6	1,517,972	6535.1	6538.5
3.4.2	Trend	2	-3265.9	1,513,886	6535.9	6536.5
3.4.3	Seasonality	12	-3254.3	1,417,513	6532.6	6573.8
3.4.4	Seasonality (only months 4,5,8)	4	-3258.6	1,435,382	6525.2	6538.9
3.4.5	Trend-seasonality	13	-3253.7	1,402,521	6533.4	6577.9

TABLE. 17 SEVERITY MODEL COMPARISON

As we specified in the theoretical preliminaries, the best among these five models is the very one that minimizes AIC and/or BIC and/or has the smallest model error in terms of quality of prediction:

Apparently, the "Gamma-seasonality (only months 4, 5, 8)" model results markedly the best between them according to AIC.

On the other hand, BIC leads us to the "Gamma-trend" model.

If we refer to the model error, the "Gamma-trend-seasonality" model has the smallest error term, and thus the best quality of prediction.

As mentioned previously, we want a model that not only has a good predictability, but also incorporates claims data features under the form of significant covariates. This means that the model must represent a sort of equilibrium between these two goals. Let us review each of these three models above and decide which severity model we will keep.

- The "Gamma-trend" model

It is the best in terms of BIC, but not in terms of AIC. Certainly, BIC criterion introduces a penalty term higher than AIC, when we encounter models with a high number of parameters, but for a model as the "Gamma-trend" model with 2 parameters, the

difference of penalty term weight between AIC and BIC is negligible. We can't quite decide with these criteria.

The "Gamma-trend" model has the second highest model error, which means a poorer quality of prediction compared to 3 out of the other 4 severity models.

Finally, its trend covariate is insignificant: With respect to a Wald test, we accepted the null hypothesis with a 2nd error type β.

Given these last two conclusions, we decide not to choose this model.

- The "Gamma-trend-seasonality" model:

Based on the model error, this model has the best quality of prediction among the five severity models. Nevertheless, it has the biggest BIC, which is explained by the high number of variables. One ought to add the fact that most of this model's covariates were not significant with respect to a Wald test. These two reasons push us not to choose this severity model.

- The "Gamma-seasonality (only months 4, 5, 8)" model

This model was chosen in terms of AIC. However, the difference of AIC values remain very small, which seems to make it a poor decision criterion in this context. A propos of its BIC ranking, it is smaller than the "Gamma-trend-seasonality" model, and has almost no difference with the "Gamma-trend" model. And finally, its model error is average, since it is smaller than the "Gamma-trend" model, but a bit higher than the "Gamma-trend-seasonality" model. One also ought to remember that all covariates of the "Gamma-seasonality (only months 4, 5, 8)" model were significant, which is an advantage compared to the other two.

In order to decide between all three models, the fact of having significant seasonal covariates only convinces us to choose the **"Gamma-seasonality (only months 4, 5, 8)"** model. It responds to our equilibrium in terms of quality of prediction, and it has the smallest AIC.

Interestingly, the "Gamma-seasonality (only months 4, 5, 8)" model considers the same average severity for months other than April, May and August, i.e., a fractioned seasonality. It remains an accurate model choice, since the inclusion of seasonality has been empirically highlighted several times in this thesis.

Conclusions:

This chapter constitutes the second part of our work regarding the hazard component, where we gauged the damage severity element and designed its model with seasonal features found in the observations.

In parentheses, we found that a time trend appears to be statistically insignificant in this severity model. On the other hand, it may be quite intriguing that a time trend covariate

in the frequency model chosen in Chapter 2 was significant in terms of a Wald test. We understand that frequency and severity components may incorporate different features, i.e., Frequency with {seasonality, trend} vs Severity with {fractioned seasonality, no trend}.

In spite of its statistical significance, we admit that the "Gamma-seasonality (only months 4, 5, 8)" model cannot predict values that reach outliers in the claims data. Remember that we evoked the possibility of treating claims separately: attritional vs extreme. As we have a look at Figure 29, one could separate approximately average claims greater and less than $3 M. On one side, attritional claims can be modeled as we already did with the "Gamma-seasonality (only months 4, 5, 8)" model. On the other side, one would use Extreme Value Theory to model extreme claims. In the case of the count data, it would be more difficult to follow the same separation approach, since the count data have less observations in number than the claims data.

Conclusively, we are thus aware of the fact that the retained models from Chapter 2 and 3 are not yet very adapted and inclusive of all features found in the observed data. The relative figures of the fitted values show that high variability above average under the form of outliers remains still outside the prediction ability of our frequency/severity models. This duo constitutes thus a basic Frequency-Severity model. We resolve eventually the problematic issue of outliers in Chapter 4 of Part II, in which we will explain and develop the Multifractal model, as it is inclusive of all variability features in general: trend, seasonality and jump outliers.

Conclusion

Basic Frequency x Severity model

Previously in Chapter 2, we developed a frequency model, while in Chapter 3, the severity model was constructed. The mixture of these two chosen models will generate estimated aggregate costs for the independent Frequency-Severity model. We will call it the Basic model.

It suffices to calculate the following quantities with the simulated data:

$$\widehat{C_t} = (\widehat{N_t - N_{t-1}}) * \hat{\hat{C}}_t$$

For example, at time t=130 that corresponds to October 2000: For the Basic model, the simulated aggregate claim will be $29.49 M.

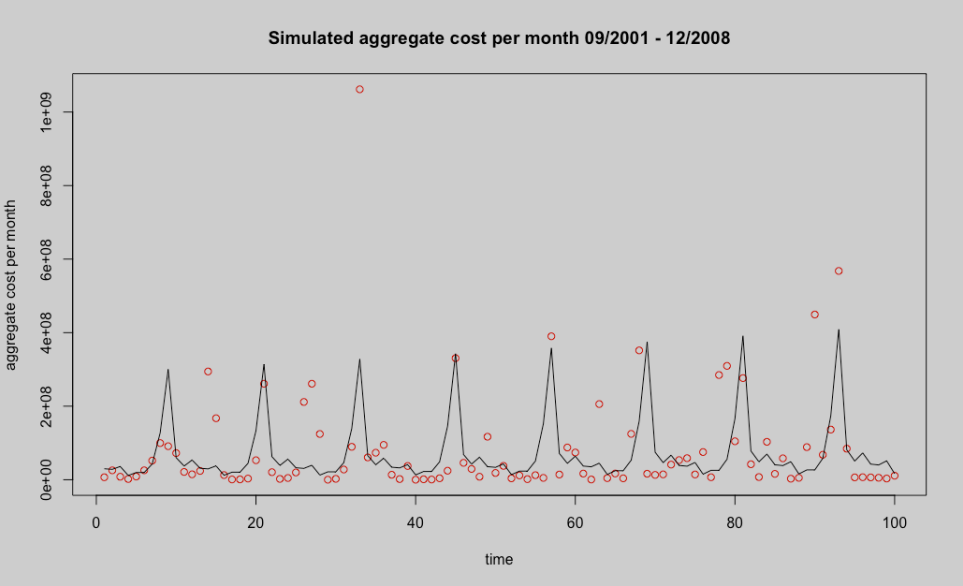

Fig. 31 Fitted values of Basic Frequency-Severity model vs the observed tornado aggregate data in red

These simulated data will be used in Chapter 5 for application purposes and will be compared to the Multifractal simulated data.

We finally calculate the error term for this model. This will help us have a clear view of its quality of prediction. The error term is calculated:

Basic Model: $Error\ term = 119.9\ M$

Part II. Frequency-Severity Cat models with Multifractals

4. Multifractals – an innovative way

In Chapter 1, we presented a real picture of natural catastrophes in the US, and more particularly of tornadoes. We then purported to model the two main components of tornado risk by the mean of frequency and severity schemas in Part I. The initial approaches lead us to some very important conclusions: a seasonality and a slightly upward trend need to be implemented in a way or another, as they measure quite well the behavior of tornado risk. In Chapter 2, a frequency model accounting for an annual seasonality and time trend was chosen. The same happened in Chapter 3: the severity model considers the feature of seasonality after a careful study.

A last but not least feature has been left out of the undertaken studies till now: the very high variability of the data. The variability above average, both for the tornado counts and monthly average claims, becomes an undeniable issue for our further study. The need for a new model arises. It should include at first seasonality and trend features to be at least as good as the former selected models in Chapters 2 and 3, and it should also add a new feature representing the high variability of the data. In this chapter, we develop the multifractals, since they accomplish variability inclusion in three levels and improve the tornado risk modeling with some of their features.

Opposite to some commercialized Cat models, the Multifractal model presented hereby doesn't take into account exposure and vulnerability. As we work on an aggregate basis (the U.S.) and exposure and vulnerability features on a local basis, one need not worry at the time being. The model in question poses fully on a probabilistic schema. One can refer to section Conclusions, Limits and Perspectives, if need be for more insight on this comparison and how the Multifractal model overcomes this obstacle.

4.1. New features

As it was mentioned in the introduction of this thesis, tens of scientific fields have interwoven methods and mathematical tools. Practical and efficient applications have been put in place, and the Multifractal model proposed by Hainaut & Boucher [4] belongs to this vast class. Our goal here is to replicate a modified version of the Multifractal model [4]. First applied to data in geophysics, the Multifractal model was merged with a classical Poisson-Gamma model regarding the tornado data. The multifractal structure was then added to the claim arrival intensity and the claim size average. Several trials took place until the best fitted model was chosen. Before schematizing the multifractal model, we will first develop a theoretical overview of multifractal processes, namely the multifractal formalism, and show why we are convinced that the current application is very efficient compared to former traditional ways of modeling non-life insurance Nat Cat data.

First of all, multifractals aim the volatility in observed data. As this volatility is not deterministic and changes with time, we call it stochastic volatility. For example, Leövey

gives us a specific insight about the importance of employing a stochastic volatility (SV) model in high frequency data [1]. Starting from basic models that include quadratic variation as a measure of volatility, to the coherent stochastic volatility models group where multifractals find their place, the main idea is related to a long-range/short-range dependence and time-varying volatility. Furthermore, "the more present the large positive autocorrelations in volatility are, the more motivated one is to use more complicated versions of stochastic volatility models" [1]. To sum up, we are interested in the time-varying and stochastic volatility components in such models. These terms will be defined later.

One would therefore ask why one ought to be interested in multifractal models. It is said to be by far one of the most important alternatives to classical stochastic volatility approaches. It belongs to a "new" generation of models, even though a classic in geophysics or other scientific fields, which is characterized by its "temporal multi-scaling and the capacity to replicate sudden increases in fluctuation spikes without recurring to boundary parameter values" [1]. Replicating specific features of time series data is made possible by the structure of multifractals. Such a model also belongs to the literature of regime-switching and Itô diffusions [2]. Multifractals are, among others, well known for their self-similarity property and their impact on Long Range Dependence (LRD) that we define in the following section. The last one is of great interest to us, given the particular volatility of the tornado data. Additionally, they dominate the other SV models at longer horizons when forecasting is studied [3].

However, multifractals are quite different compared to SV models, particularly in the construction of their basic principles. In the next section, we will not only be giving further explanations about all the new notions related to multifractals, but also will show their construction. In such a way, we can have a clear idea of how they work and their advantages. Then, we will apply this model to the Frequency and Severity components of tornado risk.

4.2. Theoretical overview

Before understanding what a multifractal is, one may ask what a fractal does stand for. They are characterized by certain irregularities occurring in data. For instance, certain real world signals as in geophysical data often possess "an erratically changing oscillatory behavior", which have earned them the name multifractals [33]. The appearance of these phenomena may make stationarity of data questionable. This is in our advantage, because the data were shown to be non-stationary in the last two chapters. For the record, data are said to be stationary if their theoretical process, say, $X(t)$, does not depend on time. The trend, another feature of them, sometimes can be explained by strong correlations (Long Range Dependence). The sudden jumps, a typical for multifractals, also represent a last feature in particular.

After having detailed some basic definitions and illustrations related to them, we will have a clearer vision of how multifractals work and which their advantages are.

4.2.1. Definitions and features

In theory, a fractal process Y (t) has a non-integer degree of differentiability, which naturally leads to an analysis of its local Hölder exponent $H(t)$ [33]. For a random polynomial P and a random constant C, the equation of the proxy is as follows:

$$|Y(t') - P(t)| \simeq C|t' - t|^{H(t)}$$

This polynomial may also simply be sometimes the Taylor polynomial of Y at t [33]. The proxy depends basically on the corresponding time interval adjusted by the Hölder exponent and a constant C.

Multifractals are nonetheless but a generalization of the fractals. Before defining the multifractals and the corresponding measure, we should shed light on another essential feature leading to a better comprehension of them.

Multifractals are characterized by a temporal time-scaling. It was Mandelbrot who claimed in first place that the "time scale selected to analyze a time series should not modify the statistical properties of the process itself" [34]. This assures us of the seasonal and monthly time scale we have chosen. This feature is highlighted by the self-similarity it incorporates:

Self-similar process (page 36 of [1]): A random process X(t) that satisfies

$$\{X(ct_1), \dots, X(ct_k)\} \stackrel{\text{def}}{=} \{c^H X(t_1), \dots, X(t_k)\}$$

for some $H > 0$ and all $c, k, t_1, \dots, t_k \geq 0$, is called self-similar or self-affine. The number H is the self-similarity index or the scaling exponent of the process $X(t)$.

While looking carefully at the condition a random process $X(t)$ must satisfy in order to be self-similar, the H exponent, as in the case of the Hölder exponent, plays here a similar role. This is a mere connection between fractal processes and the self-similarity statistical property. Also, when one mentions self-similarity, one has in mind its close link to the long-term dependence (LRD) concept. Having the LRD property means that the auto-correlation of a random process $r(k)$ decays so slowly that the $\sum_k r(k) = \infty$ [1].

Anyway, there are some computational problems that one would encounter if one did not make this definition "more pragmatic". Its theoretical representation is the obstacle here. It was, once again, Mandelbrot, who, by implementing a probabilistic model for the distribution of energy dissipation in intermittent turbulence, generalized the self-similarity paradigm. In order to avoid any kind of empirical "violations", multifractal formalism replaces the c^H of the definition above by a positive random factor M_c [35]. This replacement renders the model non-stationary.

The measure of multifractal processes makes use of this feature:

Multifractal measure (page 38 of [1]): A random measure μ defined on [0, 1] is called multifractal if it satisfies for all $q \in \mathcal{Q}$:

$$\mathbb{E}[\mu[t, t + \Delta t]^q] \sim c(q)(\Delta t)^{\tau(q)+1} \text{ as } \Delta t \to 0$$

where $\mathcal{Q}$ is an interval containing [0, 1], $\tau(q)$ and $c(q)$ are deterministic functions defined on $\mathcal{Q}$, and the operator $\sim$ implies that if $h(\Delta t) \sim g(\Delta t)$, the two functions h and g satisfy $h(\Delta t)/g(\Delta t) \to 1$.

In appendix B, an example of Mandelbrot's cascade measures, which is a basic example of multifractal measures, explains their functionality. It ought to be noted that their probability mass is conserved from one cascade to another. Also called hierarchical cascades of combinatorial nature, they constitute the first generation of models known for satisfying the above definition. Cascade measures will serve us to construct the Multifractal model.

Now that the multifractal measure is well defined, exemplified and illustrated in Appendix B, we can finally define the multifractal process, the core of the Multifractal model:

Multifractal process (page 41 of [1]): A stochastic process $X(t)$ is called multifractal if it has stationary increments $\delta_{\Delta t}X(t) = X(t + \Delta t) - X(t)$ that satisfy the moment scaling rule:

$$\mathbb{E}[|\delta_{\Delta t}X(t)|^q] \sim c_X(q)(\Delta t)^{\tau_X(q)+1} \text{ as } \Delta t \to 0$$

The multifractal process, as the fractal, basically depends on the time interval Δt with some modifications!

One would think that there is a contradiction with what was said above, when one reads this definition. The tornado data are non-stationary, while the definition of the multifractal process requires stationary increments. However, there are no doubts when we consider the following: As it was discussed above about the self-similarity, Mandelbrot replaced c^H by a positive random factor M_c. This implies non-stationary data. Thus, the probability law of $\delta_{\Delta t}X(t)$ does depend on t, and the scaling function $\tau_X(q)$ is non-linear [33]. This leads us to a multifractal process requiring non-stationary increments, as it is the case of the tornado data.

This concludes the last crucial point to be visited before turning to the Multifractal model. The multifractal formalism shown till now allows for seasonality via its time-scaling, trend via its Long Term Dependence and jumps inclusion in the next sections. We will now give the main arguments how this formalism is appropriate to our case and why we find it advantageous compared to the other methods mentioned in the beginning of this chapter. This will be followed by the multifractal model applied to the tornado data. In the end, a numerical application will be furnished.

4.3. Applying multifractals to our schema

The previous theoretical overview lists some of the reasons why multifractals are a very useful tool when it comes to modeling count time series that have features such as a trend in time and sudden jumps. Also, they allow us to model a random process that has some memory of past information in the data. As it was shown in Chapters 2 and 3, monthly GLM approach with time trend alone was not sufficient to generate fitted values that can reach the observed extreme data showing their high volatility. The observed data are said indeed to be of a regime switching-frequency variation, but we weren't accounting for all regimes in the frequency/severity models. In order to overcome this obstacle, we found that multifractals can imply all regimes.

Why all this interest on multifractals, one would simply ask? Because a single mechanism "can capture changes in regimes of low frequency variations, can specify intermediate frequency dynamics known for smooth autoregressive processes, and can finally capture high frequency switches" [3] in form of jumps here. In other terms, a new alternative model involving regime-switching is proposed, whatever the frequency form. In this last one, the regime switches at an infrequent rate at predetermined dates, which turns the model to a non-stationary one.

Let us also recall that the data involve climatic phenomena. On one hand, climatic factors affect the frequency. On the other hand, social and economic factors affect the claim severity, as for example the inflation, the degree of industrialism of the affected zone, population density, etc. It is very complicated to take all these factors into account in the Multifractal model as it is almost impossible to collect the data. For instance, the causes of the vortex inside a tornado or how a tornado will arise are still unknown by meteorologists. In such a case, we consider these factors, whether they're climatic or social/economic, as unobserved. Anyway, these last ones can be involved indirectly, via a reconstruction of the existent Multifractal model.

We used a generalization of the multifractal process to model the unobserved factors of the count time series [4,36]. This is made obvious by the very interpretation of this model, whereas "the underlying process can represent such latent unobserved factors" [5]. The volatility is simply represented by the product of a finite number of random multipliers - the latent unobserved factors. Following Calvet & Fisher work, parsimoniously, these multipliers are supposed to be 1st order Markov and identical except for time-scale, as they vary with time. Finally, we assume the multipliers have identical marginal distributions and differ only in their switching probabilities. This model "delivers long-memory features in volatility, substantial outliers, and a decomposition into components with heterogeneous decay rates" [36].

Till now, multifractals were mainly used in a financial time series framework [36], where the Gaussian assumption remains unchanged. In finance, the Gaussian distribution property has been improved by introducing "more realistic volatility features" [1]. As it turned out though, financial times series, like several other time series arising from physics, astronomy, biology and medicine, are not really self-similar but resemble the aggregative normality feature. They have indeed thinner tails and become less peaked in

the tails when the sampling interval decreases [1]. The approach of multifractals is being exploited for the first time in a non-Gaussian framework, i.e., in the tornado data [4].

In the next section, the Poisson-Multifractal model on the tornado arrivals will be constructed by following a step-by-step method. The binomial multifractal process is employed on this purpose [3]. The Gamma multifractal model follows after.

4.3.1. Frequency with multifractals

The Poisson-Multifractal model can account for over-dispersion and time dependence at the same time [5]. By taking the classical Poisson distribution, we want to add the past information. The same notations as in [4,5] will be employed from now on for verification and comparison purposes. Remember that N_t is the number of tornadoes that have occurred during a period of time t. N_t is a process defined on a filtration, noted F_t, in a probability set (Ω, P). In the Probability theory, $F_t = \sigma(\Omega)$, a sigma algebra of Ω, contains the past information about the history of N_t. λ_t, the intensity of N_t, is a stochastic process too, and is defined on a filtration H_t, carrying on information about the history of λ_t alone. $H_t \vee F_0$ notation contains the initial filtration F_0 augmented by the filtration of λ_t. λ_t is said to be constant during Δt, the length of a period. Putting all the information altogether, N_t, conditionally on $H_t \vee F_0$, is a Poisson process with the following probability formula when n jumps have been observed:

$$\Pr(N_t = n | H_t \vee F_0) = \frac{\left(\sum_{i=0}^{t} \lambda_i \Delta t\right)^n e^{-\sum_{i=0}^{t} \lambda_i \Delta t}}{n!},$$

It is a doubly stochastic process, also called the Cox process [37]. λ_t, the mean intensity, is the product of a piecewise constant function $\lambda(t)$ and a multifractal process F_t^N.

Remark: To verify the fractal nature of F_t^N by a visual analysis of auto-covariance in levels, please refer to [4].

We denote by λ_t the stochastic intensity on $[t, t+1)$:

$$\lambda_t = \lambda(t) F_t^N = \exp(\beta^T x_t) F_t^N,$$

where $\lambda(t) = \lambda_i$, $i = t mod 12$, the vector x_t contains data covariates, and F_t^N represents the multifractal process, which is a a time-dependent function. $\lambda(t)$ is a constant piecewise function representing empirical average of the monthly tornado counts.

Remark: If we also had an exposure, as it is the case in most of non-life insurance data, we would simply multiply the expression above by the exposure variable.

In F_t^N we introduce then m^N random factors as a multiplicative product. Again, they're unobservable and are being modeled by a first-order Markov finite state vector, which

leads to an easier construction of the covariates over time, and, as it will further be shown, allows for the ML estimation:

$$\bar{M}_t^N = (M_{1,t}^N, M_{2,t}^N, \dots, M_{m^N,t}^N) \in \mathbb{R}_+^{m^N},$$

and so the multifractal process obeys the following form:

$$F_t^N = \prod_{j=1}^{m^N} M_{j,t}^N.$$

The volatility components $M_{j,t}^N$ are, among others, persistent, positive-valued and satisfy the following average condition $\mathbb{E}(M_{j,t}^N) = 1$ (see Appendix 2 for more details). The last condition assures us that at least λ_t equals $\lambda(t)$ in average [4]. By simplicity, the random factors at a given time t are statistically independent [3].

$M_{j,t}^N$ is built in a recursive way by using Bayesian updating procedure. All the random factors follow an identical process, except for time-scaling variation. For $j \in \{1, \dots, m^N\}$,

$$M_{j,t}^N = \begin{cases} M_{j,t-1}^N \; with \; 1 - \gamma_j^N \\ M^N \; with \; \gamma_j^N \end{cases},$$

where M^N represents a fixed distribution:

$$M^N = \begin{cases} m_{0,j}^N \; with \; p_0 = 1/2 \\ 2 - m_{0,j}^N \; with \; p_0 = 1/2' \end{cases}$$

which shows the binomial nature of the multiplier, as M^N takes equiprobable fixed values from $\{m_{0,j}^N, 2 - m_{0,j}^N\}$. If $M_{j,t}^N$ (indirectly M^N) receives $2 - m_{0,j}^N$, the regime switches and increases the intensity. γ_j^N is namely the high-frequency probability that $M_{j,t}^N$ switches.

F_t^N can take $d = 2^{m^N}$ values, which will be noted $s_1^N, \dots, s_d^N$. Given that M^N has a discrete distribution, it will only generate a finite number of volatility states. Each of these values is obtained by a multiplicative product of $M_{j,t}^N$, $t = 1, \dots, T$. Given the nature of the vector $\bar{M}_t^N$, we generate F_t^N as a Markov chain, whose dynamics are featured by the following transition matrix:

$$A^N := (a_{i,j}^N)_{1 \le i,j \le d},$$

where $a_{x,y}^N = \Pr(\bar{M}_t^N = s_x^N \,|\, \bar{M}_{t-1}^N = s_y^N) = \prod_{j=1}^{m^N} \left(\gamma_j^N \frac{1}{2} + (1 - \gamma_j^N) \mathbb{I}_{(M_{j,t}^N = M_{j,t-1}^N)}\right)$.
($\mathbb{I}_{(M_{j,t}^N = M_{j,t-1}^N)}$ is an indicator function)

The difficulty arises when one needs to know the values of $M_{j,t}^N$, which are unobservable at first. However, the solution has already been introduced in the literature of

multifractals. Following what was explained on [5], where the authors employ a filtering technique developed by Hamilton [38] and inspired by Kalman's filter [39], we can generate probabilities of being in a state given all the observations in the past.

Let us denote by n_i, $i \in \{0, 1, \dots, t-1\}$, the total number of tornadoes observed in the precedent period, given an observed number of tornadoes n_t at time t. The likelihood function is indeed a vector, including probability functions per each value of the multifractal process, say, F_t^N:

$$p(t, n_t, x_t) = \begin{pmatrix} \Pr\left(N_t = n_t | \bar{M}_t^N = s_1^N, x_t\right) \\ \vdots \\ \Pr\left(N_t = n_t | \bar{M}_t^N = s_d^N, x_t\right) \end{pmatrix}.$$

For example, for $x_t = j$, the probability distribution takes the following value:

$$p(t, n_t, j) = \Pr\left(N_{t+h} - N_t = n_t \middle| \lambda_t^j\right) = \frac{(\lambda_t^j \Delta t)^n e^{-\lambda_t^j \Delta t}}{n!}$$

Furthermore, we need the probability of being in a certain state j, which is written as follows:

$$\Pi_t^{(j)^N} = \Pr\left(\bar{M}_t^N = s_j^N | n_1, \dots, n_t, x_t\right).$$

Now, we will make use of Hamilton filter to calculate $\boldsymbol{\Pi}_t^N = \Pi_t^{(j)^N}{}_{j=1,\dots,d}$ recursively as a function of past probabilities.

$$\boldsymbol{\Pi}_t^N = \frac{p(t, n_t, x_t) * (\boldsymbol{\Pi}_{t-1}^N A^N)}{\langle p(t, n_t, x_t) * (\boldsymbol{\Pi}_{t-1}^N A^N),\ \mathbf{I} \rangle'}$$

where $\mathbf{I} = (1, \dots, 1) \in \mathbb{R}^d$ and $x * y$ indicates the Hadamard product $\begin{pmatrix} x_1 y_1 \\ \vdots \\ x_d y_d \end{pmatrix}$.

One last condition must be satisfied before we explain the recursion method. Parsimoniously, we assume that the Markov processes in the Poisson-Multifractal process have reached their stable distribution. As Calvet & Fisher have specified [3], this is quite common when we are dealing with empirical applications. In such a case, $\boldsymbol{\Pi}_0^N \sim$ Ergodic Distribution, which is characterized by the eigenvector of the initial matrix A^N related to the eigenvalue equaling 1. Given that we have assumed the independence of the m^N random factors $M_{1,1}^N, M_{2,1}^N, \dots, M_{m^N,1}^N$ at time $t = 1$, the atoms of $\boldsymbol{\Pi}_0^N$ are uniquely determined by:

$$\Pi_0^{(j)^N} = \prod_{l=1}^{m^N} \Pr\left(M^N = s_{0,l}^{j\ N}\right), \text{ for all } j = 1, \dots, d.$$

Let us finally introduce the log-likelihood, conditionally on the observed count data till time T:

$$\ln L(n_1, \dots, n_T \left| m^N_{0,j=1,\dots,m^N};\ \gamma^N_{j=1,\dots,m^N} \right.) = \sum_{t=1}^{T} \ln \langle p(t, n_t, x_t)\ , (\mathbf{\Pi}^N_{t-1} A^N) \rangle.$$

In order to obtain the maximum likelihood estimator - MLE, we apply a numerical maximization to the previous log-likelihood equation. The maximization problem is stated as follows:

$$\max_{\hat{\theta}} \ln L(n_1, \dots, n_T | \hat{\theta})$$

where $\hat{\theta} = \{m^N_{0,j=1,\dots,m^N};\ \gamma^N_{j=1,\dots,m^N}\}$ is the desired MLE.

For m^N being a finite scalar, and $T \to \infty$, the MLE is consistent and asymptotically efficient.

4.3.2. Severity with multifractals

In Chapter 3, we retained a Gamma distribution to model the claims data. The idea is quite similar: the Gamma-Multifractal model follows the same construction path as Poisson-Multifractal's. Multifractals will be involved in the mean parameter of costs. We denote C_t the claims cost caused by one tornado having lieu in $[t, t + \Delta t)$. As demonstrated along the thesis, we verified that there are certain trends, namely, stronger seasonality for one period than for another, and pics in April-May, showing thus very high variability in the cost data. These trends are to be included via multifractals.

C_t, as N_t, is defined on the filtration F_t. The mean cost of C_t is denoted by τ_t. $E_t = \sigma(\Omega)$, the filtration on which τ_t is defined, carries on the information about the history of τ_t alone. Then the distribution function of costs per tornado C_t having occurred in $[t, t + \Delta t)$, conditionally on $E_t \vee F_0$, is a Gamma process:

$$f(c|E_t \vee F_0) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu c}{\tau_t}\right)^{\nu} \exp\left(-\frac{c\nu}{\tau_t}\right) \frac{1}{c} \mathrm{I}_{(0,\infty)},\ \nu \in \mathbb{R}^+.$$

Remark: As in the case of Poisson-Multifractal, all formulae, except for the distribution probabilities formulae, and the notations are the same. We change only the index N to C.

The mean parameter of C_t is defined as the following product:

$$\tau_t = \tau(t) F_t^C = \exp(\beta^T x_t) F_t^C,$$

where $\tau(t) = \tau_i,\ i = tmod12,\ \tau_i$ are the monthly claims cost averages; F_t^C is the multifractal process of Gamma.

Remark: Refer to [4] for a visual analysis of auto-covariance in levels of C_t.

In this step of the construction, an important assumption should be taken into consideration; the cost severity is independent from the tornado/claims arrival. This assumption can be omitted; we have treated this issue later on. The scale of cost is influenced by m^C random factors, which are unobservable and independent from the m^N factors in the above Poisson-multifractal model. The Markov state vector is written as follows:

$$\bar{M}_t^C = (M_{1,t}^C, M_{2,t}^C, \dots, M_{m^C,t}^C) \in \mathbb{R}_+^{m^C},$$

and so the multifractal process obeys the following form:

$$F_t^C = \prod_{j=1}^{m^C} M_{j,t}^C.$$

The multiplier $M_{j,t}^C$, is drawn from a fixed distribution M^C with $\Pr = \gamma_j^C$, otherwise it equals $M_{j,t-1}^C$. They are in average equal to one: $\mathbb{E}(M_{j,t}^C) = 1$ (see Appendix 2 for more details). F_t^C is adapted to the filtration E_t and can take $d = 2^{m^C}$ values, which will be noted $s_1^C, \dots, s_d^C$. The transition matrix of the multifractal F_t^C is given by A^C. Finally, if we denote by $\mathbf{\Pi}_t^C$ the vector of probabilities, and $c_1, \dots, c_T$ the observed costs up to time T, then the likelihood is given by:

$$\ln L(c_1, \dots, c_T \,|\, m_{0,j=1,\dots,m^C}^C;\ \gamma_{j=1,\dots,m^C}^C, \nu) = \sum_{t=1}^T \ln \langle f(c_t)\,, (\mathbf{\Pi}_{t-1}^N A^C) \rangle,$$

where $f(c_t)$ denotes the density vector $\left(f(c_t|E_t \vee F_0)\right)_{j=1,\dots,d}$.

The parameter set $\{m_{0,j=1,\dots,m^C}^C;\ \gamma_{j=1,\dots,m^C}^C, \nu\}$ is to be estimated by numerical maximization of the above log-likelihood.

4.3.3. Improvements of Frequency and Severity with multifractals

They are applied to both models: The Poisson-Multifractal and Gamma-Multifractal.

Our goal is to reduce the dimension of the parameter set. The more parameters there are, the more complicated a model can be. Such models cost in terms of computation time too. Numerical maximization works well with few parameters, and it follows that the number of iterations shouldn't in general be greater than a fixed scalar. We can improve the Multifractal model by reducing the number of parameters to be estimated. Calvet & Fisher [36] have already tried by applying the following simplifications to $m_{0,j=1,\dots,m}$ and $\gamma_{j=1,\dots,m}$:

According to empirical analyses, having different $m_{0,j}$ for all j "often leads to a better fit of count data" [5]. A possible parametrization is:

$$m_{0,j} = (m_0)^{j^c}$$

We reduce the parameter set of the fixed distribution M from m parameters to only two, $\{m_0, c\}$.

Remark: An even simpler case would be $m_{0,j} = m_0$, for all j.

We also can specify the heterogeneous high-frequency transition probabilities $\gamma_j = (\gamma_1, \gamma_2, \dots, \gamma_m)$ as:

$$\gamma_j = 1 - (1 - \gamma_1)^{b^{j-1}},$$

where $\gamma_1 \in (0,1)$ and $b \in (1, \infty)$ [2]. Calvet & Fisher [2] have shown that, for small values of j, $(\gamma_1)^{b^{j-1}}$ remains small and $\gamma_j \sim (\gamma_1)^{b^{j-1}}$. As a result, "probabilities of low-frequency components grow approximately at geometric rate b". On the other hand, at higher frequencies ($\gamma_j \sim 1$), the rate of increase slows down. γ_j's condition above assures us that it is a probability measure (being less than 1). By this construction, we easily satisfy the condition:

$$\gamma_1 < \gamma_2 < \dots < \gamma_m < 1.$$

In [5], intending for the approach to capture low-valued regime shifts and long volatility cycles of count data, it is shown via the inference method that, if γ_j is inversely proportional to j, the last factor $M_{m,t}$ changes its value less frequently than the first factor $M_{1,t}$.

With such a formulation, all $\gamma_1, \gamma_2, \dots, \gamma_m$ depend on γ_1, thus reducing the set from m parameters to only two. In this framework, $\{\gamma_1, b\}$ is the new parameter set for transition probabilities.

Conclusion: We have finally a set of 4 parameters, $\{\gamma_1, b, m_0, c\}$, to estimate with the numerical maximization (also add ν for Gamma-Multifractal).

4.3.4. Two-dimensional Frequency-Severity with multifractals

This section is given for information and future study purposes; we will treat the issue of accounting for a possible correlation structure between Frequency and Severity Multifractal models, but won't give details about its construction or application. The modeling of this correlation parameter is quite difficult and somehow unrealizable in this thesis, given the restricted time framework.

One may ask about a probable relation between frequency and severity datasets. It must be stated here that a simplifying assumption of independence holds between the bi-dimensional data. It can be unrealistic though and it wouldn't provide the most accurate representation of claims amounts, as they are likely to be dependent on the claim counts. For example, a month with several tornadoes might only generate small claim amounts while a month with only one claim might in fact submit a higher-than-average claims amount. The correlation coefficient represents indeed a basic structure of the probable existing dependence. We have a weak-to-moderate empirical correlation of approximately 0.3 between the number of tornadoes and the average cost observations. This fact helped us assume that a certain dependence pattern might be useful.

The independence assumption that is held in this thesis would be dropped. Indeed, a simple way of inducing a bivariate correlation would be to consider common random factors for both models, i.e., these random factors are generated by the same multifractal process. This would be made possible by a simultaneous fitting [4], aka a Poisson-Gamma two-dimensional model, where both multifractal processes would come from the same generating parametrization. In section 4.3.3, improvements were made, and a set of 4 parameters, $\{\gamma_1, b, m_0, c\}$, was chosen to be estimated with the numerical maximization for the Poisson and Gamma separately. In a two-dimensional model, this parameter set would be the same for both (also add ν for Gamma-Multifractal).

In [4], not only some common random factors are included, but also a parameter for the unconditional correlation between the arrivals of the two multifractal processes is shown. We could refer to [4] for a detailed modeling of this parameter. This is only one way other researchers have already developed in their field [24]. Some other ways will be mentioned while showing the limits and perspectives of the Multifractal model toward a commercialized Cat model in the end of this thesis.

4.3.5. Numerical Application

4.3.5.1. Poisson-Multifractal

We may now apply the improved Multifractal model under the independence assumption to the observed data and estimate the parameter set. Different numbers of random factors will be tested and the selection criteria from Chapter 2 will help us decide which the best model is. In the end, a backtesting technique on the last 100 observations will be used to see how well the Multifractal model performs based on the fitting to the tornado data.

First, we will be dealing with the Poisson-Multifractal from section 4.3.1, with improvements from section 4.3.3, which led us to a set of 4 parameters: $\{\gamma_1, b, m_0, c\}$. In order to maximize the log-likelihood function that was presented thereby, we implemented it in SAS and used the methods discussed earlier. The code was furnished by the author of [4,5], and improved by us, for example by changing the size of the parameter set. It must also be specified that the Newton-Raphson method was used in order to maximize the log-likelihood. The SAS code takes no more than a computational

minute to furnish us with the log-likelihood corresponding to Poisson-Multifractal for a given number of random factors, denoted Nbm, and the estimated most likely parameters. For $Nbm > 7$, the optimization becomes very heavy, because of a transition matrix of more than 4^7 rows [4]. We decided to test models having up to 7 random factors.

No exact statistics exist currently to test the goodness of fit of a regime-switching model [4]. J.P. Boucher in [4] has constructed an empirical approximated chi-square statistic for this purpose. It uses the simulated intensities, $\bar{\lambda}_t$, which incorporate the average (seasonality and trend) and the volatility (multifractal component) of the Poisson-Multifractal process, and the observed tornado counts, N_t. We denote it by Z^N and calculate it as follows:

$$Z^N = \sum_{t=k}^{T} \frac{(N_t - \bar{\lambda}_t)^2}{\bar{\lambda}_t} \sim \mathcal{X}^2_{k-4}.$$

Z^N follows approximately a Chi-square distribution with $k - 4$ degrees of liberty, where 4 is the number of estimated parameters. In order not to be influenced by the estimated intensities directly dependent on the probabilities $\Pi_0^{(j)^N}$ [4], Z^N and the corresponding p-value must be calculated using the last observations of the sample, hence those of the backtesting sample of 100 observations. The smaller Z^N is, the better the prediction and the stronger the acceptation of null hypothesis will be.

In [4], it has already been shown for the tornado count data that for $Nbm \leq 5$, the empirical chi-square statistic disqualifies the corresponding models, because of a p-value less than 5%. Therefore, we will only be showing Poisson-Multifractal with $Nbm \in \{6, 7\}$, for which cases we have a p-value >> 0.05. The log-likelihood values will be compared between estimations based on the raw count data vs smoothed count data as it was done in [4]. c is the forth parameter from the improved parameter set. With the estimates of this parameter set, we can also deduce the values of the other transition probabilities and the values of the fixed distribution M from section 4.3.3. A certain stability of some of the estimated parameters can be observed in table 18 [4].

To be estimated	Nbm	J.P. Boucher model [4]	Improved model [4,5]
$Log - likelihood$ γ_1 b m_0 c $m_{0,j}, j \in \{1,2,3,4\}$ $(\gamma_1, \gamma_2, \dots, \gamma_6)$	6	−1,150.353 0.23 1.785 0.7056 -	-946.872 0.77 2.097 0.543 0.9137 0.54, 0.72, 0.54, 0.41 0.77, 0.57, 0.31, 0.09, 0.01, 0.00002
$Log - likelihood$ γ_1 b m_0 c $m_{0,j}, j \in \{1,2,3,4\}$ $(\gamma_1, \gamma_2, \dots, \gamma_7)$	7	−1,159.252 0.262 1.569 0.7448 -	-947.565 0.77 1.721 0.543 1.003 0.54, 0.65, 0.42, 0.27 0.77, 0.57, 0.31, 0.09, 0.01, 0.00, 1.931E-10

TABLE. 18 POISSON-MULTIFRACTAL COMPARISON BASED ON THE NUMBER OF RANDOM FACTORS

The estimation of transition probabilities $(\gamma_1, \gamma_2, \dots, \gamma_m)$ is based on γ_1 and b from the parameter set. This means that changing the number of random factors m won't affect the number of parameters to be estimated. We can hence compare between models with 6 or 7 random factors. The difference in terms of likelihood is almost insignificant: one could choose either between them.

We represent graphically a simulated set of tornado counts with the Poisson-Multifractal having 7 random factors. We chose a backtesting sample of 100 observations, as for the empirical statistic Z^N discussed above. The corresponding p-value equals $0.6 >> \alpha = 5\%$, so we keep the model with 7 random factors by committing an error of 2nd type β. The results seem to be quite satisfactory. The variability above average in terms of low, medium or high-frequency, say the seasonality, trend and the overall volatility, is finally taken into account by the help of multifractal components in the Poisson-Multifractal.

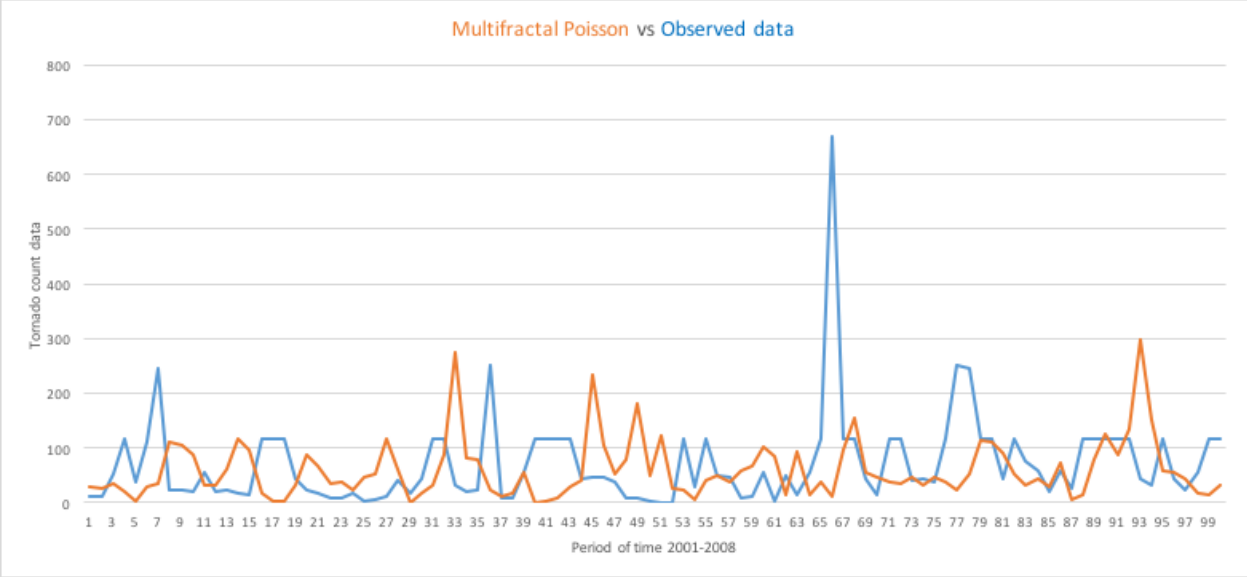

Fig. 32 Poisson-Multifractal simulated values vs backtesting observed sample

We also furnish the monthly average tornado counts based on the simulated data. They will further be used in the application in Chapter 5.

Month	Observed mean of count data	Simulated mean
January	25.53	25.16
February	24.63	32.58
March	54.53	79.21
April	85.74	99.48
May	119.68	162.28
June	76.74	110.58
July	47.63	67.64
August	32.95	32.66
September	39.84	30.15
October	37.74	50.08
November	48.16	63.82
December	15.37	13.67

TABLE. 19 OBSERVED MEAN OF TORNADO COUNT DATA VS POISSON-MULTIFRACTAL SIMULATED MEAN

Before we jump to biased conclusions regarding the backtesting technique, it is interesting to present some preliminary observations based on the results shown in Table 19 above. Remember that the tornado "season" with the highest number of tornadoes per month happens during the April-June period. The multifractal simulated data seem to amplify the respective means. For instance, regarding the month of May, the observation

shows 110 occurred tornadoes, while the simulation indicates 162. This could be due to a very frequent switching intensity of multifractal random factors, given the May observations. Thus, during the tornado "season", the random factors based on the past information switch their regime more than during the rest of the year period. Future empirical studies might give us a better insight of this amplification.

4.3.5.2. Gamma multifractal

For the Gamma-Multifractal, we follow the same path as we did with Poisson-Multifractal. We have this time a set of 5 parameters, $\{\gamma_1, b, m_0, c, \nu\}$. The empirical statistic [4] this time considers the simulated monthly average costs, denoted $\bar{\tau}_t$, and their estimated standard deviations, and is calculated as follows:

$$Z^C = \sum_{t=k}^{T} \frac{(C_t - \bar{\tau}_t)^2}{\bar{\sigma}_t{}^2} \sim \mathcal{X}^2_{k-5}.$$

Z^C follows approximately a Chi-square distribution with $k-5$ degrees of liberty, where 5 is the number of estimated parameters. As shown in [4], the p-values for whichever $Nbm \in \{4, 5, \dots, 9\}$ are near 1. The log-likelihood is quite stable and there are no significant variations when the Nbm increases. See [4] for more details. Interestingly, we saw the same phenomenon on the log-likelihood variation of severity models in Chapter 3.

We decide to keep the model with 7 random factors in order to have a similarity with the Poisson-Multifractal. The estimations from SAS output are given hereby [4]. The figure 33 shows simulated vs observed values.

$$log - likelihood = -3223.846$$
$$\widehat{\gamma_1} = 0.1646$$
$$\hat{b} = 7.3779$$
$$\widehat{m_0} = 0.6816$$
$$\hat{c} = 1.0000$$
$$\hat{\nu} = 2.5341$$

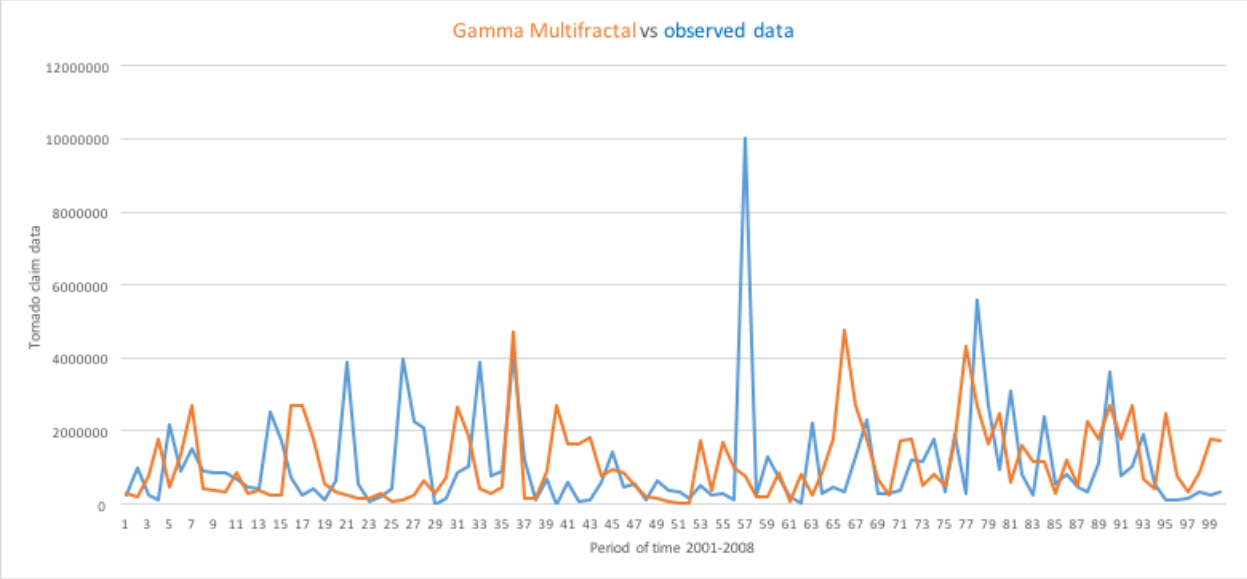

Fig. 33 Gamma multifractal simulated values vs backtesting observed sample

The monthly average tornado claims in thousands of dollars, based on the simulated data of the backtesting technique are given below. They will also be used when pricing a derivative-reinsurance instrument in the application in Chapter 5.

Month	Observed mean of claim data (in $k)	Simulated mean (in $k)
January	593	588
February	1,105	1,604
March	1,081	1,414
April	1,403	1,843
May	2,363	2,602
June	581	572
July	484	492
August	1,531	1,944
September	460	517
October	762	953
November	1,071	1,254
December	492	250

TABLE. 20 OBSERVED MEAN OF TORNADO CLAIM DATA VS POISSON-MULTIFRACTAL SIMULATED MEAN

We can clearly see how the simulated data follow closely the observed cost intensity per each month. Again, as it was the case with the Poisson-Multifractal, the Gamma-Multifractal seems to switch their regime in order to have a similar pattern with the observed data, and they also appear to amplify them. May and August values represent the most affected months by tornado claims, and the multifractals take that well into

account. For the record, a similar amplification was observed with the Poisson-Multifractal too. Future studies might show us how and why this amplification effect happens.

We finally calculate the error term for the Multifractal model (Poisson x Gamma) in order to opine on its quality of prediction when compared to the Frequency-Severity model from Chapters 2 & 3. The error term is calculated:

Multifractal: $Error\ term = 82.732\ M$

Conclusion: In Chapters 2 and 3, frequency and severity models were constructed. The seasonal and trend components were taken into account, and two models were kept after having applied the selection criteria. However, the seasonal component, alone or along with the trend component, was shown to be insufficient to capture finer behaviors. The multifractal model appears to overpass this obstacle. The multifractal model is indeed almost complete. The frequency and severity components have been modeled with a seasonal component and a certain number of random factors were added, which helped us to include variability of low, medium and high frequencies in terms of volatility. Now, seasonality, trend and some jump outliers can be fitted reasonably well. However, the highest peak is still unattainable in both tornado arrivals and tornado claims. An improvement would be to cap it and see how this affects the modeling results. It was also shown that the multifractals can also amplify the observed average data. This could be part of a future research on a similar topic.

Regarding the Gamma-Multifractal, one can bear in mind the fact that the improvement in terms of likelihood is not very significant. As well as for the severity model in Chapter 3, we can tell the same thing. We also understand that the inclusion of random factors in this severity model with multifractals needs to go under other tests than an empirical Chi-square statistic.

Finally, regarding the backtesting figures 32 and 33, the volatility component of multifractals allows us to follow a similar pattern to the observed data pattern. Anyway, some jump outliers were left unattainable. It is interesting to see that we have an amplification of the simulated means in Tables 19 and 20 on one hand, and some jump outliers still left outside the prediction on the other hand. To sum up, given the conclusions from the empirical statistics, the multifractal model is certainly a very good model for what it was discussed above. We still expect future studies to show how the multifractals can be improved when it comes to a more complete inclusion of the observed jump outliers. A less strong amplification would be another goal of the studies in question.

For what it follows in Chapter 5, Poisson and Gamma multifractals will be compared with the models from Chapters 2 and 3. A derivative-reinsurance instrument will be developed in the following chapter, and it will serve to compare between the two frameworks, a basic Frequency-Severity model vs the Multifractal model. We will see how the volatility added by the multifractals gives more accurate price rates.

5. Derivative-reinsurance instruments – a numerical application

Tornadoes appear to be a very frequent natural catastrophe that can generate huge amounts of damage costs. When accumulated, these property claims surpass billions of dollars to be indemnified. If we use the Fujita scale for measuring them, we would mostly be interested in tornadoes belonging to an F2 or higher category, as they are the most disastrous tornadoes. In Chapter 1, we observed the overall impact of severe thunderstorms during a twenty-year period. It was further shown that approximately 40% of the insured losses are caused by events including tornadoes. In 2015 dollars, we would count for $162.5 billion of losses. What was even more surprising about tornado facts and figures is the increasing trend of their arrivals. More and more tornadoes are happening each year. In the same plurennial study, we may distinguish the yearly amounts of losses corresponding to this upward trend. It is more than obvious that this natural catastrophe represents a great risk for the trio of decision-makers in Chapter 1: insurers, reinsurers and the government. The trend in question has to be taken into account in the pricing of whichever financial/ (re)insurance product. We will briefly resume the issue in its most basic form.

Insurers sell insurance coverages that protect insureds against losses caused by Nat Cat, such as tornadoes. In most of the cases, damages become unbearable for an insurance company. This will at first affect its "losses to premiums" ratio, and then decrease the reserves aiming these losses. Insurance companies operate locally most of the time. So, they do not necessarily diversify such risks. If a whole area is prone to tornadoes, there is a big chance of that area to be hit by a tornado. In the end, the shareholders' capitals and the capacity to underwrite more coverage will be reduced. It is thus that uninsured losses come into life and the gap between insured losses and economic losses (insured + uninsured) gets bigger. And it is up to government funds, if any, to cover the gap. Insurers have to cover themselves by the help of reinsurers/investors.

Reinsurers provide insurance for the insurance companies. How does this work? In definition, "the reinsurer agrees to indemnify the cedent (the insurer) against all or part of a loss which the ceding company may incur under certain policies of insurance that it has issued. In turn, the cedent pays a consideration, typically a premium" [40]. It is exactly the case of insurance companies that have portfolios of policies "heavily exposed to catastrophic events and which have a strong need for reinsurance cover" [40]. Several advantages are known for the insurance company, such as capital relief, or a better underwriting policy. Reinsurers also have an accurate expertise in managing Nat Cat risks, as their portfolios cover policies spread nationwide. By diversifying such risks, both the insurance and reinsurance companies have less need of required capital. Nat Cat are also said to be the main driver of reinsurance [41].

Our goal is to measure the last element of tornado risk: insured risk. Now that we have a quick insight of mechanics behind the way how (re)insurance companies deal with Cat risks, we reach the conclusion that a (re)insurance company will certainly cover its claims

in the reinsurance market or in the derivatives' market by the mean of reinsurance treaties, derivative instruments, or both. In the following sections, we will first discuss in general the existing reinsurance treaties, whether they are traditional or nontraditional, and give details especially about those which may be of interest to us. Afterwards, we will present some derivatives' notions and formulae that we found helpful for our study. Given the form of the observed data, we choose then to conceptualize a product that has characteristics from reinsurance treaties and weather derivatives.

5.1. Traditional reinsurance treaties

One may find two basic types of reinsurance agreements: facultative and treaty reinsurances.

- Facultative reinsurance is based on an individual risk basis. The ceding company has the option to offer an individual risk to the reinsurer and the reinsurer retains the right to accept or reject the risk. [42] It is less adapted for Nat Cat coverage, because the reinsurer may decide not to cover the policies with bad/very bad results as when very severe tornadoes occur. This would certainly lead to less coverage of Nat Cat due to their severity, and consequently to more uninsured losses caused by them. From the economic point of view, this would increase the insurance protection gap.

- Treaty reinsurance is a transaction encompassing a block of the ceding company's book of business, or a part of portfolio. The reinsurer must accept all business included within the terms of the contract [42]. It is preferred in the case of Nat Cat, given the trajectory of a tornado affecting partly/wholly a specific zone. If a zone is considered as one portfolio, for example the Dallas zone, then it is more interesting to cover a part or the totality of policies in that zone. This helps in diversifying the risk.

Given the corresponding arguments above, we decide to continue with the second type of arrangement, the treaty reinsurance characteristics. Regarding the concept of loss that we will be reinsuring, the treaty can cover a specific risk/policy, or an event, i.e., all policies affected by it. It will then be called per-risk or per-event/occurrence according to one's choice. Nat Cat is normally covered by a per-event type of treaty reinsurance.

One would also find two main categories in which one could dissect the treaty reinsurances: pro rata and excess of loss. This categorization shows the way how losses and premiums of the cedent are split between the cedent and the reinsurer.

- A pro rata, aka "proportional", shows which percentage of premiums and losses is ceded to the reinsurer. It is usually associated with a limit per event in the case of Nat Cat. "Pro rata forms are often used in property insurance, since this form provides catastrophic protection in addition to individual risk capacity" [42].

- An excess of loss, aka a "non-proportional" treaty, denoted XOL, is related to the term of "retention". "The reinsurer does not get involved with a loss until a predetermined retained limit of loss or retention, which the ceding company will pay, is exceeded" [42]. Unlike a pro rata, the premium ceded to the reinsurer is not proportional to the premium income the ceding company has collected from the insureds. It is quite useful when it comes to "underwriting large risks as a protection against severity of loss" [42].

The standard method for Nat Cat losses in the reinsurance market is the excess of loss. It is what we call a "Catastrophe per-event Excess of Loss" (abbreviated as Cat XOL). "Its purpose is to protect the ceding company against loss experience resulting from the accumulation of losses arising from a single, major natural disaster or event such as a tornado. For a given event, the treaty applies once the accumulation of losses are paid by the reinsurance company, less the predetermined retention" [42]. We will now focus on this treaty and detail its characteristics.

5.1.1. Catastrophe per-event Excess of Loss - Cat XOL

Let us give a brief theoretical overview of a Cat XOL with a predetermined priority and how it is applied to claims data. The illustration from [42] will help us understand the functioning of this treaty. As mentioned previously, the retention must be fixed in advance in order to define which part of the ceded claims will be covered.

The illustration proceeds: An insurer needs some extra capacity to write a property business of $1 M for competition purposes. The company decides to transfer a part of its portfolio under an XOL. The study that it undertakes determines the retention, aka priority, of $300 M. It basically means that, for some given event claims, it can retain up to $300 M of that amount. If the amount of event claims is less than the retention, the insurer reimburses the whole event claims. If the event claims exceed the retention, the exceeding portion of the amount will be covered partly or wholly by the reinsurer. For the part of event claims higher than $300 M, reinsurance must be applied. For instance, an XOL would be covering up to a specific retained limit = $700 M for the ceded event claims that exceed the retention of $300 M. The XOL treaty is written under the form: 700 M XS 300 M. The limit simply is the sum of the retention and the retained limit.

Remark: In some cases, the treaty may be unlimited; the reinsurer may cover unlimited sums exceeding the retention.

To sum up the charge to be covered by the reinsurer, for claim denoted by C, retained limit by L, retention by R, and a limit of U:

$$Reinsurer's\ charge = \begin{cases} 0 & if\ C < R \\ C - R & if\ R < C < U \\ L & if\ C > U \end{cases} = \min\left(\max(0; C - R)\,; L\right)$$

And with the example data, we would have:

$$Reinsurer's\ charge = \begin{cases} 0 & if\ C < 300\ M \\ C - 300\ M & if\ 300\ M < C < 1.000\ M \\ 700\ M & if\ C > 1.000\ M \end{cases} = \min\left(\max(0; C - 300\ M)\,; 700\ M\right)$$

For a claim of \$600 M: The insurer pays \$300 M, and the other \$300 M will be under the reinsurer's charge.

For a claim of \$1.200 M: the reinsurer covers only \$700 M; the part of \$200 M exceeding the limit of \$1.000 M and the retention of \$300 M will be covered by the insurer.

The excess of loss is valuated as it is usually done with the insurance products; a commercial premium rate is fixed and needs to be payed to the reinsurer. Techniques for reinsurance treaty valuation are different, as they can be deterministic or stochastic. If the observed claim data is used for the pricing, the deterministic method will be applied. The calculation of the rate in the second case is said to be stochastic, as the data is not observed but simulated with a probabilistic method. We call it the statistical method.

5.2. Nontraditional reinsurance treaties

5.2.1. Insurance Linked Securities (ILS) context

(Re)insurers may still have difficulties in managing all risks they have undertaken by underwriting traditional treaties. It is thus more than necessary for them to transfer some part of their portfolios to other reinsurers/market investors/government by means of different techniques. The first one would be the retrocession, which is a kind of transfer where the reinsurer cedes a percentage of his portfolio to another reinsurer. It is a quite common technique.

A new technique, which has been developed in the late 90', is about the securitization of the (re)insurance portfolio. They are called weather derivatives. R.L. McDonald defines a financial derivative as "a financial instrument - agreement between two moral people - that has a value determined by the price of an underlying asset", and a weather derivative as "a derivative contract with a payment based on a weather-related measurement such as heating or cooling degree days" [43]. Weather can be very unpredictable, which makes a weather derivative a risky instrument and difficult for the pricing.

The securitization method has been employed by reinsurers such as Swiss Re in the case of Cat bonds covering hurricane claims. [4] explains how this contract works in its paper: "The reinsurance treaties are transferred to a special purpose vehicle company, and in exchange for collateral, investors receive a periodic floating payment, linked to the amount of claims covered by treaties". For a detailed version of Cat bonds application related to tornado claims, one might refer to [4], and study how this technique is used under a multifractal formalism framework. In [4] the authors have shown the efficacy of such a method. We won't be developing any further, given the large choice of research

papers on this topic: Cat bonds at [4], weather derivatives at [44], flood bonds at [45], Cat bonds with credit risk at [46], etc.

To sum up, reinsurers and/or investors may propose treaties or ILS products to insurers in order to help them extra-cover their portfolio risks. It is up to the insurer to evaluate the productivity of each instrument and how it affects its solvency ratios. We are presenting a new derivative instrument in the following section. Investors would propose it to insurers in order to cover their event claims. This instrument would use the backtesting results in annual losses of tornado arrivals from Chapters 2-3, and 4, and will provide us with a price of insured risk.

5.2.2. Our proposal: "Asian Cat XOL"

If the costs of each individual event (tornado) were available, the application of a Cat XOL in Section 5.1.1 and the calculation of the reinsurer's charge would consequently be direct. We have monthly aggregate costs of cat events in the database instead. Anyway, this obstacle is handled quite well if we introduce derivatives into our study. We will be indeed conceptualizing a new instrument, called "Cat XOL with an Asian Call component", which is similar to a Cat XOL, but has an Asian Call option component in it. We will refer to it as *"Asian Cat XOL"*. This would be able with the following:

- One would certainly see something very similar between an XOL reinsurer's charge and an Asian Call option with average price.
- The valuation technique becomes applicable by using a Monte Carlo method, as one would do for the stochastic pricing of a Cat XOL.
- Given its particular form, we will also be able to take into account the seasonality in the pricing of this special derivative treaty.

We will be dealing with these points in the following sections.

5.2.2.1. Use of Asian Call options

In order to clarify the similarity between an XOL reinsurer's charge and an Asian Call option with average price, we will briefly discuss Asian Call options with average price and their pricing method. The definitions of Call options and Asian Call options are needed first.

Definition: Call Option (page 62 of [43])

A Call option is a contract where the buyer has the right, but not the obligation, to buy an underlying asset S (for ex: stock, claim) for a fixed price K at a predetermined expiration date T.

If the price of the option at T is greater than K, the buyer exercises the option and buys the underlying asset at K, and earns the difference between price of the underlying asset at T, say, S_T, and K. The payoff is written as follows:

$$\text{Payoff}_\text{T} = \begin{cases} 0 \text{ if } \text{S}_\text{T} < \text{K} \\ \text{S}_\text{T} - \text{K} \text{ if } \text{S}_\text{T} > \text{K} \end{cases} = \max(\text{S}_\text{T} - \text{K}; 0)$$

Definition: Asian Call option (page 475 of [43])

An Asian Call option has a payoff that is based on the average price of the underlying asset, which we denote as $\bar{S}_T$ over some period of time $\{0, \dots, T\}$. It is a kind of path-dependent option, which means that the value of the option at expiration T depends upon the path by which the underlying asset arrived at its final price. Its payoff has the following representation:

$$\text{Payoff}_\text{T} = \begin{cases} 0 \text{ if } \bar{S}_T < \text{K} \\ \bar{S}_T - \text{K} \text{ if } \bar{S}_T > \text{K} \end{cases} = \max(\bar{S}_T - \text{K}; 0)$$

We are interested in an Asian Call option with a maturity of 1 year and an arithmetic average price:

$$\bar{S}_T = \frac{\sum_{i=0}^{T} S_i}{T}$$

How would one calculate this average? Given the particular structure of such a path-dependent option, we consider paths of the underlying asset price and calculate the arithmetic average for each path. In our case, stocks are replaced by the aggregate claims. A path of stock price would right now be a path of aggregate claims. In respect to the seasonal feature, we will state 12 paths, where each path signature is related to one month of the year. For the path of January, for instance, we will only consider January aggregate claims during that period.

5.2.2.2. Construction of "Asian Cat XOL"

The derivative treaty that we will consider would need the following changes on the underlying asset and exercise price, where $\bar{C}_T$ stands for the monthly average cost severity in $[0, T]$:

$$\bar{S}_T \Rightarrow \bar{C}_T$$
$$K \Rightarrow R$$

We consider an unlimited Cat XOL Aggregate treaty to simplify our case. We are now able to define:

Cat XOL with an Asian Call component

The Asian Cat XOL has the following form:

$$Payoff_j = \begin{cases} 0 & if\ \bar{C}_T < R \\ \bar{C}_T - R & if\ \bar{C}_T > R \end{cases} = \max(\bar{C}_T - \mathrm{K}; 0)$$

Where j represent one of the twelve monthly paths.

It has more or less the same characteristics as a Cat XOL and an Asian Call:

- Its payoff is calculated as the reinsurer's charge in Section 5.1.1. The buyer of the treaty, i.e., the insurer, has the obligation to pay the monthly aggregate claims that do not exceed the retention and the seller, aka the investor, those that exceed that threshold.
- It is similar to the Asian Call because of its arithmetic average price component and the twelve monthly paths.

Interestingly, the reinsurer's charge and the Asian Call option payoff have the same formulation. This allows us to aim the well-known method of Monte Carlo.

5.2.2.3. Functioning of "Asian Cat XOL"

How would this instrument work in reality? The Asian Cat XOL would be traded in an ILS market, for instance. The potential investor would sell the Asian Cat XOL to the insurer. The underlying "asset" would be calculated on the US portfolio of the aggregate claims caused by tornadoes. The retention R is determined as the threshold that exceeds the attritional claims in order to eliminate the last ones. The insurer would pay to the investor the price calculated with a Monte Carlo method, denoted by P here. Per each month, if the monthly aggregate claim exceeded the deductible, the investor would pay to the insurer the difference for that month. We say that the insurer would "buy" the claims at a fixed amount R. In another case, it would pay nothing but the calculated price P, and the claim charge comes back to the insurer. The P transaction here would be equivalent to a reinsurance premium transfer from the insurer to a reinsurer.

The following schema would facilitate the understanding of the Asian Cat XOL:

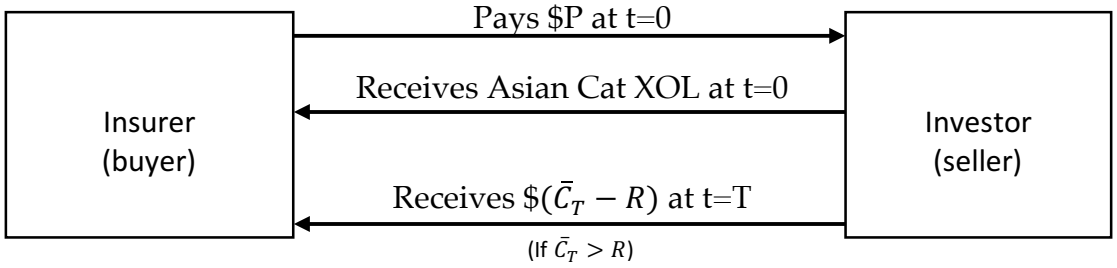

Fig. 34 Schema of an Asian Cat XOL transactions

		Payoff at t=T	
Insurer	*t=0*	$\bar{C}_T < R$	$\bar{C}_T > R$
Buys 1 treaty	$-P$	0	$\bar{C}_T - R$
Net total		$\boldsymbol{-P}$	$\boldsymbol{\bar{C}_T - R - P}$

TABLE. 21 INSURER'S PAYOFF FOR AN ASIAN CAT XOL

5.2.2.4. Monte Carlo pricing method

For pricing purposes, we will employ the Monte Carlo method, which is commonly employed when one deals with Asian Calls. For the backtesting period 09/2001-12/2008 as in Chapter 4, we have a hundred data of monthly aggregate claims. In them, there are 8 or 9 elements per monthly path.

So, for month j and backtesting period $\{0, \dots T\} = \{Sept\ 2000, \dots, Dec\ 2008\}$, and based on the corresponding monthly path, the monthly average aggregate claim is calculated with the Monte Carlo principle based on the simulated data:

$$\widehat{\bar{C}_T^{\,J}} = \frac{\sum_{i=0}^{T} \widehat{C_i^{J}}}{T}$$

And so, the payoff for monthly path j:

$$\widehat{Payoff_J} = \max\left(\widehat{\bar{C}_T^{\,J}} - R;\ 0\right)$$

The final result under the form of a price of this instrument will be the average of Monte Carlo paths of the payoff:

$$\widehat{Price} = \frac{\sum_{j=1}^{12} \widehat{Payoff_J}}{12} = \frac{\sum_{j=1}^{12} \left(\max\left(\widehat{\bar{C}_T^{\,J}} - R;\ 0\right)\right)}{12}$$

In order to express the Asian Cat XOL as a rate of the earned premium income, denoted EPI, on the US property portfolio of the present year, one would calculate this as follows:

$$\widehat{Rate} = \frac{\widehat{Price}}{EPI}$$

5.2.3. Application of "Asian Cat XOL"

5.2.3.1. Inputs and parameters

Three elements should be revised before we go further with the application:

- The simulated aggregate claims

It suffices to calculate the following quantities with the simulated data:

$$\widehat{C_t} = \widehat{(N_t - N_{t-1})} * \hat{\bar{C}}_t$$

For example, at time t=130 that corresponds to October 2000: For the basic chosen Frequency-Severity model in Part I, the simulated aggregate claim will be $29.425 M, and for the multifractal data $25.908 M. The observed data give a real aggregate claim of $25.495 M.

- Retention of Asian Cat XOL

In the case of the retention threshold, multiple comparable studies tackle the issue. We want to find the extreme claims exceeding the threshold. The claims up to the threshold are considered as attritional claims. In the reinsurance field, the cedants will specify a retention that is relative to their underwriting policy and their risk appetite. In the Extreme Value Theory, the Peaks-Over-Threshold - POT - method tries to distinguish between attritional and extreme peaks in the data.

In order to simplify this issue, the retention will equal the first quartile of 25% given by the boxplot of the aggregate data: a retention of $27.96 M. A future study may be held in order to improve the Asian Cat XOL retention and its results.

- Earned premium income - EPI

We should finally have an EPI that corresponds to the US property portfolio. For this, we refer to Insurance Information Institute data. According to their calculations, the estimated earned premiums from the US Property portfolio were around $150 billion in 2001. Also, the percentage of losses caused by tornadoes was around 40%. So an easy estimate would be 40% of $150 billion, say $60 billion, would represent the US property EPI for the tornado coverage.

5.2.3.2. Numerical application

Now, we can apply the Asian Cat XOL with the calculation steps from section 5.2.2.4 to the simulated data from the basic chosen model of Frequency-Severity from Chapters 2 and 3 and to the multifractal data from Chapter 4 for the backtesting period. This will give us the accuracy of both simulated data and show how much improvement in terms of stochastic volatility the Multifractal model brings based on the price of the Asian Cat XOL. We start with the monthly aggregate claims: table 22 gives aggregate claims of the basic model and the multifractal model for the beginning 12-month of the backtesting period. Then, table 23 gives the calculated payoffs. The price and rate of the Asian Cat XOL are given afterwards in table 24.

Monthly path	Basic Model (M $)	Multifractal (M $)
Jan	15.30	16.43
Feb	119.87	203.66
Mar	43.55	70.31
Apr	36.24	58.06
May	55.42	36.10
Jun	90.04	124.47
Jul	29.12	36.95
Aug	41.86	70.56
Sep	26.84	17.32
Oct	25.42	53.03
Nov	32.44	88.95
Dec	10.35	3.80

TABLE. 22 MONTE CARLO PATHS: BASIC MODEL VS. MULTIFRACTAL

Monthly path	Basic Model (M $)	Multifractal (M $)
Jan	0	0
Feb	92.02	125.70
Mar	15.70	22.35
Apr	8.39	20.09
May	27.57	8.14
Jun	62.19	96.51
Jul	1.27	8.99
Aug	14.01	22.60
Sep	0	0
Oct	0	0
Nov	4.59	6.99
Dec	0	0

TABLE. 23 PAYOFF: BASIC MODEL VS. MULTIFRACTAL

Results	Basic Model (M $)	Multifractal (M $)
Price of Asian Cat XOL	18.81	25.87
Rate	0.031%	0.043%

TABLE. 24 PRICING RESULTS: BASIC MODEL VS. MULTIFRACTAL

In order to compare between the two models in terms of pricing, a prediction error was calculated in the previous chapters. We had the following results:

Multifractal: $Error\ term = 82.7\ M$
Basic model: $Error\ term = 119.9\ M$

This helps us conclude that, based on the model error, the Multifractal model has a better quality of prediction than the Basic model.

Given that this instrument is in its earliest stages, we do not possess a market benchmark which would allow us to have an insight about its "true" price. One would think that comparing with current traded Cat bonds or other ILS securities would give a general idea. However, we do not intend to start a new study on ILS benchmarks, so we found that it is better to refer to a comparison of the quality of prediction. A simple prediction error statistic as the one we just calculated is a simple method to decide between both models and conclude which one valuates more fairly the price of the Asian Cat XOL.

Based on our results and the model error calculation previously, we see that the rate of 0.031% from the Basic model seems to be underestimated. This is mainly explained from the fact that only seasonal and trend features are taken into account. The high variability of the aggregate data is partially attained in this model. On the other hand, the Multifractal model seems to be priced at a fairer rate, since it considers more volatility components than the Basic model. In Chapter 4, we studied this model and added 7 random factors to it. We purported to include variability and the results were representative of the real aggregate claims. The difference between priced rates, approximately at 0.012%, shows, among others, how a model with a dynamical component appears to be better than a model with a seasonal/trend component. If an investor were to propose such an instrument, it would be on its best interest not to underestimate the possible charges coming from tornado claims, as they appear to be very volatile along the time. In the simple case that we considered in the beginning of this section, the investor would better sell the Asian Cat XOL at a price of 0.043% of the overall EPI.

Conclusion:

Reinsurance treaties, whether they are traditional or not, must be priced prudently, especially when they cover very risky natural perils such as tornadoes. As tornadoes tend to become more and more frequent, the insurer knows that it is very probable it will reimburse claims caused by them in the future. This risk also has its own "season" of the year, being from April to end of June, when most of the tornadoes occur. So the pricing method of a such an instrument must incorporate the seasonality and high variability of these claims.

A pros/cons table as given below shows some advantages and disadvantages of the two models:

Pros/cons	**Basic Model**	**Multifractal**
Seasonality (medium volatility)	+	+
Trend (high volatility)	+	+
Outliers (high volatility)	+/-	+/-
Low volatility	+	+
Number of parameters	-	+
Prediction error	-	+

TABLE 25. PROS AND CONS BY COMPARING BASIC MODEL & MULTIFRACTAL

In [4,44,45,46], Cat bonds covering different perils have been priced and a volatility component was always present. In this Chapter, we proposed a new derivative-reinsurance instrument that also did consider the volatility component. It followed the structure of a Cat XOL, introduced features of an Asian Call, and it was priced with the Monte Carlo method. The rate over the EPI was given in the end. We do not pretend that Asian Cat XOL is better or worse than Cat bonds or any other treaty covering natural perils. Keep in mind that we were aiming an instrument that remains competitive when it comes to modeling such a seasonal and volatile risk as tornadoes. Future studies could improve this instrument by adding other finer features in order to make it more attractive for an investor in the ILS market.

Conclusion, Limits, and Perspectives

"All models are wrong, but some are useful", said George E.P. Box, a British statistician. In any case, this statement shouldn't discourage us from developing models and measuring their predictability and performance. We are aware of their limits and should have them in mind for future improvements, mathematically or conceptually speaking.

In Chapter 4, we already listed some improvements in order to reduce the dimension of parameter set. It was costly in terms of computational time to keep the former version as shown in [4]. With a set of 4 parameters, the SAS code took less than 3-5 minutes to run and provide its outputs even for 7 or 8 random factors. Also, in [4], a two-dimensional Multifractal model was studied, having not only common random factors, but also a variable that counts for dependency between them. Unfortunately, no more than 6 common random factors could be tested for this two-dimensional model. When the number is higher than or equal to 7, a transition matrix of more than 4^7 rows [4] would be needed! Even though this last model was far too complex for our thesis, the results from [4] are promising for future research.

As we go through the last chapters, the initial model went under some quite drastic changes: from a pure Frequency-Severity model, namely, Poisson-Gamma here, to the Multifractal model that is statistical and has a stochastic component too. During this thesis modeling chronology, i.e., Chapters 2 to 4, some limits were eliminated by adding first seasonality, then time trend, and at last variability with random factors. Other limits and perspectives were pointed out as well. However, one would ask if one can find other limits/perspectives by comparing with another Cat model. For a moment, let us assume no other statistical model can reach the Multifractal model complexity. One might find it as a strong reason to conclude that this model remains the best up-to-date. Certainly not! Opposite to statistical models, there is the class of dynamical models. They have proved their efficacy and are widely employed by reinsurance companies in the pricing of treaties on Nat Cat. We refer to them as Vendor models. RMS, AIR, and EQE Cat are some of the market leaders in commercializing such models.

In the following discussion, information from guidelines or presentations of current Vendor models was reviewed. Without breaching the confidentiality, a specific study of one commercial Vendor model on convective storms such as tornadoes [48] was mostly referred. The details of the methodology they use won't be mentioned, as this information remains confidential. This comparison was crucial to state the Multifractal model in front of Vendor models.

Vendor models vs multifractal model

It was mentioned that there is a huge difference between two major classes: statistical and dynamical models. Statistical models are very good at analyzing historical data,

evaluating trends and extrapolation [48]. However, they appear to be unable to capture more complex processes. During our thesis development, this was quite obvious. On the other hand, dynamical models, broadly used in meteorology for weather forecasting and climate modeling [48], are very high-resolution and complex tools that allow the risk management to model such a complex risk, i.e., tornadoes. They simulate the laws of physics, the drivers behind the genesis of tornado risk. In between, the Multifractal model stands. It is interesting to finger out the fact that we use a statistical model that has a dynamical component, the random factors. The fact of it being a mixture leads us to an easier comprehension of its advantages and disadvantages.

Let us first recall the schema of risk evaluation from Chapter 1. We stated that the risk would regroup four main components; the hazard, the exposure, the vulnerability and finally what will be insured, say, insured risk. In our thesis, hazard was covered through Chapters 2 to 4, and insured risk by Chapter 5 with commercial applications. Vendor models also take into account the total amount of values present in the tornado-prone area, and the potential damages, say, the exposure and the vulnerability respectively. This schema of risk evaluation will help us compare between Multifractal and Vendor models. Limits and perspectives will be shown per each component.

- Hazard

It is the first component, which generates the event set and its severity. An essential input is the historical data; both statistical and dynamical models start from this point. Regarding the definition of a tornado event for example, the Multifractal model converges with Vendor models. Even though certain events are not considered to be catastrophic, because they only generate attritional losses, they are still included in the dataset, as they highly contribute to the annual aggregate loss. Very severe events are not capped but are taken entirely; they are important outliers and incorporate variability and the peak-reaching capability of tornadoes. We also verified that the tornado dataset comes from very accurate sources, the same that the leading Vendor model refers to when updating its database [48]. The U.S. National Weather Service is one among others. It goes back to 1950, and it uses the Fujita scale to categorize tornadoes. The Vendor finally uses dynamical model output "in the form of re-analysis data" [48]; it considers all information available in real time. The Multifractal model should be updated with new information too in order to be an accurate representation of the reality and avoid biased simulations.

Tornado path and scale damage

We encounter two potential limits that the Multifractal model might have towards the Vendor model. The first one is about the event footprint, and the other about the distinction of tornadoes according to the Fujita scale classification. In the reports of tornadoes in the Vendor's database, the footprint of a tornado has a beginning point or a beginning coordinate and its end point, creating a tornado line path of a specified length [48]. These line paths are very realistic when simulated with high-resolution dynamical models. We do not refer directly to a tornado line path, but only to the number of tornadoes. Also, it is crucial to distinguish tornadoes of an F1 scale from those with an F2 or higher scale. One could go even further by differentiating a tornado into these two portions: the geographical part that can be run through from a tornado that has an F1

kind of damage, and the other one with an F2 or greater kind of damage. For tornadoes reaching a 200 mile per hour wind speed, a particular distinction would be necessary too.

We have somehow resolved the problem by implementing the dynamical component of multifractals. Regarding tornado footprints, remember that multifractals represent these unobserved factors, such as climatic ones. One would say that the way a footprint is designed is directed by the wind speed, which is nothing but one climatic factor in itself. We assume this is included in the number of random factors. Also, as far as we are concerned about the scale of damage intensity along and across the footprint, the unobserved factors in the severity model take into account partially the intensity of a tornado damage. One potential improvement may be put into perspective here; we could add a specific parameter that allows differentiating between tornadoes regarding the severity scale, for example.

Tornado generating process:

This is another point where Multifractal and Vendor models converge, with the exception of a single assumption. In both cases, the generating process of tornadoes is simulated with a highly complex mathematical model. The way the parameters are defined makes the difference between the two models. Remember that, for a tornado to occur, some meteorological conditions should be met, as it was said in section 1.2.2. Most of them are not yet known; tornado genesis still constitutes an advancing scientific field.

The Vendor model assumes that "the physics of convective processes driving tornadoes are well understood" [48]. It considers thus what is called a hybrid approach. In its parameter set, two parameters are defined as sufficient given the science advances the energy needed to create the vortex, namely convective available potential energy - CAPE, and the shear, basically the force that provides rotating updrafts by producing a mesocyclone [48]. Based on thousands of combinations of the simulated values of duplets, tornadic events are simulated. Seasonality is thus taken into account. However, thunderstorm activity is not a certainty under those circumstances. The hybrid approach uses the duplets as a proxy. With this duplets, cartography of tornadoes can be generated in the time period to be considered.

The Multifractal model doesn't assume a good understanding of the generating process. We still lack a considerable comprehension of the physics driving a tornado. We prefer to keep these unobserved and use random factors to include them. In our case, the CAPE and shear parameters of the Vendor model would only be two out of random factor set. We saw that the Multifractal model can result in very satisfactory outputs up to 7 or 8 random factors. The other random factors would stand for other unknown/unobserved factors. This makes the Multifractal model more accurate than the Vendor model. The hybrid approach of the last one has definitely a limit. With the science advances, more will be discovered about tornado genesis and more generating parameters will be implemented in order to have a more realistic generating process.

Seasonal feature:

It was taken into account by both models.

Continental scale:

The development of dynamical convective event models is still very much an advancing field of science. It is not feasible at this stage in their development to use these models to generate events on a continental scale [48]. They should still find a way out to include correlation between risk portfolios of different geographical zones. On the other hand, the Multifractal model works on a continental scale, the U.S. territory, which shows another of its advantages.

Computational time:

Vendor models are usually very complex and need a great amount of time to run and give their results. On the other side, the Multifractal model takes less than 5 minutes.

- Exposure and vulnerability

The Vendor model uses these two other components to evaluate the tornado severity. Given the hazard set in a specific geographical zone and its probability of occurrence, the total insured sum of the property objects in the zone and their vulnerability to damages from a tornado, and finally the damage Fujita scale of a tornado, the simulated damage losses will make it able to evaluate the derivative-reinsurance instruments. How does it work? For each location, we have simulated values of the convective indicators CAPE and shear throughout each simulated day. This joint duplet determines the probability of an event occurring at that location [48]. For a sample set of events that we will sample from all locations, they will be weighted by their probability. The higher the exposure and the weight in those locations, the bigger the damage will be. However, despite the science advances, dynamical models still have significant biases, and their results need to be interpreted carefully.

On the other hand, the Multifractal model is only based on the historical data of direct property and crop losses. Random factors, which are here considered unobserved too, may also represent social and economic factors. As one may read in Chapter 1, we saw that social and economic factors are closely related to the exposure and its vulnerability to tornado damages. They will help us generate events of different damage intensities. There can be some limits we are currently not aware of. We saw that the improvement in terms of likelihood seemed to be not very significant when the number of random factors increased from one model to another. Future research may see how these random factors affect a severity model and how they may be related to exposure and vulnerability components. Then, improvements can be put in perspective.

In order to resume what was shown above, the following table gives the pros and cons of Vendor and Multifractal models. We used the conclusions of each feature in order to decide if a model has an advantage or not related to the feature in question.

	Vendor model	**Multifractal**
Tornado data	+	+
Tornado damage scale	+	+/-
Tornado generating process	-	+
Seasonality and trend	+	+
Continental scale	-	+
Computational time	-	+
Exposure and Vulnerability	+	+/-

TABLE 26. PROS AND CONS BY COMPARING VENDOR MODEL & MULTIFRACTAL

- Insured risk

This final component is easily described while one applies derivative-reinsurance instruments by the help of a Vendor or Multifractal model. Data must be converted into the current year US dollars, so that the direct losses taken into account are comparable on an annual basis. Given the simulated sets from both models, the treaties will be valuated and the desired net price will result in the end of the computation.

With the Vendor model, the event set can reach tens of thousands of simulated tornadoes. For instance, a set of 65,000 events is created [48]. These events will be distributed in the entire geographical area of the covered US zone. Each one has a probability of weight that determines the chances of that event to occur during the time period. With all the components discussed above, different financial and statistical indices are given under the form of a summary. Net Pure Premium associated with a standard deviation, among others, is furnished by the tool. In order to perform calculations with a Vendor model, exposure and vulnerability data for the US zone, especially property sums insured of the US portfolio, are absolutely necessary. Unfortunately, we couldn't find this information and try to use the Vendor tool.

Regarding the Multifractal model, by the help of techniques discussed in Chapter 5, the simulated data was used to valuate a derivative-reinsurance instrument such as an Asian Cat XOL, or a Cat bond as in the case of [4,44,45,46]. The application covered the US zone. From a commercial point of view, a company that wants to buy such an instrument needs to be cautious when interpreting such results. The outputs from the Multifractal model are to be compared with the results of Vendor models afterwards. It is up to its level of performance indicators and its solvency policy that it will decide which model suits best. The market benchmark rates also are a very important reference.

Conclusions: The final model retained in this thesis, namely, the Multifractal model with random factors, gave quite satisfactory results; the derivative-reinsurance instrument application was a mere proof of it. We first started from a basic model for frequency and severity components, and then improved it by adding seasonal and trend features into it. But this was not sufficient if one deals with the predictability issue. The seasonal and trend components via GLM covariates couldn't capture even finer behaviors such as variability above average that the data incorporate. The vast range of applied statistical models, even though commonly used in the non-life (re)insurance, did not convince us either. In [4], the Multifractal model, initially used in geophysics [34], and later in financial pricing [1], was applied to non-life insurance data with great success. This was one of the main incentives that pushed us to demonstrate plentifully the utility of the Multifractal model in this thesis. Adding random factors to the GLM statistical model Poisson-Gamma was a very strong "add-in": this states the Multifractal model to a higher stage compared to existing models.

Should we appreciate the Multifractal model that much, it was not a reason to show that it has a few limits too. In a comparison with a very competitive class of the so-called dynamical models, we also tried to show which potential limits it has and the performance of its risk schema evaluation in some points. Vendor models came to our help, as they are the main resultants of dynamical models so far known. By having listed some advantages and disadvantages compared to Vendor models [48] above, we were able to see which unknown limits may come over the surface and if the Multifractal model implies perspectives of improvement for Vendor models too. Consequently, we clearly stated the pros and cons of multifractals vis-à-vis the dynamical approach.

In Chapter 4, we applied the Multifractal model to the tornadic data and followed a rigorous path by adding the aforementioned features. One might ask if one could use this framework for other natural catastrophe risks. There certainly exist other risks of the same kind that could incorporate the same features. We have the floods, for instance, which are seasonal and may have a certain trend. There is the risk of winter storms in Europe, too. An empirical research would consider these risks and present the results, which we would wish to be as good as they appeared to be in the case of tornadoes. One would also want to know if there is a market for the instrument "Asian Cat XOL" in Chapter 5. It represents a marginalized topic in this thesis, as the volatility features were the main discussion topics. However, we were able to show the efficacy of the Multifractal model with this instrument, especially the minimization of the prediction error compared to the Basic Model from Chapters 2 and 3 and the inclusion of the seasonal feature via the simulation technique. Given the "infantry" of the Asian Cat XOL, we are aware of its limits and realistic pricing. We expect future studies to study the robustness of the pricing techniques and the construction of similar benchmarks.

Multifractals could potentially lead the way of studying the actual climate trend. It was proved successfully in [4]. We think we made some improvements to their model in this thesis too by proving its trend features. Now, it is up to the main decision players to implement this powerful mathematical tool in their risk-based frameworks and test its efficacy.

Appendix A

Hypotheses: Frequency * Severity = All Costs

All costs represent the aggregate claims S_i per each month. We assume independence between the frequency and the severity. In the case of the tornado data, the model assumptions are that the tornado count time series follow a Poisson distribution while the jumps, which represent the cost sizes, follow a gamma distribution.

We will give the mathematical proof of the following [24]:

"Modeling the means of the individual claim amounts $\boldsymbol{Y_{ij}}$ is equivalent to modeling the mean of the average severity $\boldsymbol{\bar{Y}_i}$ given $\boldsymbol{N_i}$".

Proof:

Consider the aggregate claims on the individual level. For month i, we have that the overall cost is $S_i = \sum_{j=1}^{N_i} Y_{ij}$, where N_i counts the tornadoes, Y_{ij} are the individual claim amounts (unknown here), i.i.d., given N_i, for $j = 1, \ldots, N_i$.

Let $\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}$ be the average claim size, or severity, with $\bar{Y}_i = 0$ when $N_i = 0$. This average severity clearly depends on N_i. Then the aggregate claims can $S_i = N_i * \bar{Y}_i$. Thus we have that the aggregate cost is the product of the claim frequency and severity. If it is further assumed that the claim frequency is independent of the individual claim amounts, a restrictive assumption, then the GLM structure for the aggregate claims is simplified. In this setting, we have that N is independent of $\bar{Y}_i$. If we assumed that at the level of monthly data, the claim amounts Y_{ij} are i.i.d. with $Y_{ij} \triangleq Y_i$, then we can write the mean aggregate claim amount as:

$$E(S_i) = E\big(E(S_i|N_i)\big) = E\left(E\left(\sum_{j=1}^{N_i} Y_{ij}\,\middle|N_i\right)\right) = E\left(\sum_{j=1}^{N_i} E(Y_{ij}\,|N_i)\right) = E\left(\sum_{j=1}^{N_i} E\big(Y_{ij}\big)\right)$$
$$= E\left(N_i * E\big(Y_{ij}\big)\right) = E(N_i) * E(Y_{ij})$$

Equivalently, in terms of the average claim severity $\bar{Y}_i$, we have that the mean aggregate claim cost can be written as:

$$E(S_i) = E\big(E(S_i|N_i)\big) = E\left(E\left(\sum_{j=1}^{N_i} Y_{ij}\,\middle|N_i\right)\right) = E(E(N_i * \bar{Y}_i|N_i)) = E(N_i * E(\bar{Y}_i|N_i))$$
$$= E(N_i) * E(\bar{Y}_i|N_i)$$

Since the mean claim cost is the product of the mean frequency and mean severity in the independent model, then in a GLM framework, the model for S_i is simply the product of the marginal GLMs for N_i and $\bar{Y}_i$ given N_i respectively.

It follows that modeling the means of the individual claim amounts Y_{ij} is equivalent to modeling the mean of the average severity $\bar{Y}_i$ given N_i.

End of the proof!

Appendix B

We present Mandelbrot's cascade measures hereby, taken explicitly by [47]. The construction of these cascades starts by assigning a random measure μ_0 to a bounded interval [0, 1], say, and consequently employing an iterative transformation to it. In the first step, μ_0 is divided into two subintervals receiving positive constants m_0 and m_1, respectively. In this simple version, the constants may be chosen to obey $m_1 = 1 - m_0$, with $0 \leq m_0 \leq 1$. The resulting measure is commonly referred to as μ_1, which is no longer uniform in [0, 1] but rather has a step-function shape; the left interval has a height of m_0, while the right one has a height of $1 - m_0$. In the next cascade step, the two intervals of μ_1 are split up again into two subintervals receiving factors m_0 and $1 - m_0$, assigned from left to right. This leads to the measure μ_2 consisting of four intervals, each with its own probability mass: ${m_0}^2$, $m_0 * (1 - m_0)$, $(1 - m_0) * m_0$, $(1 - m_0)^2$. The procedure can be repeated ad infinitum leading μ_1 to weakly converges to the measure μ.

One should note that these transformations never alter the total mass of μ_0, they only spread it off by the factors m_0 and $1 - m_0$ along the original support. One speaks in this case of a conservative measure, given that the original mass is preserved at each iteration step. It follows that in any interval of size $\Delta t = 2^{-k}$, the probability mass amounts to:

$$\mu_k[t, t + \Delta t] = {m_0}^{kv_0}(1 - m_0)^{kv_1},$$

where $t = \sum_{i=1}^{k} \eta_i 2^{-i}, \eta_1, \dots, \eta_k \in \{0, 1\}$, and v_0 and v_1 denote the relative frequency of 0s and 1s in the series $(\eta_1, \dots, \eta_k)$ [47]. Also, the nature of the factor assignment to each interval of size 2^{-k} has led μ to be referred to as the Binomial Multifractal measure.

In order to illustrate how the distribution evolves from one cascade to another, the following construction is taken explicitly from [1]. This is an example of the construction of a conservative measure with random variables $M \in \{m_0, (1 - m_0)\}$ and $m_0 = 0.7$. From top to bottom: a draw of the first level μ_1, a draw of the second level μ_2, a draw of the 12th level μ_{12}, and an empirical sample series of squared returns of the Japanese Yen (YEN) as a proxy for volatility. The latter series consists of ≈ 16.25 years of data starting on the 2nd of January of 1979. For better visualization, μ_k was multiplied at each cascade level k by 2k.

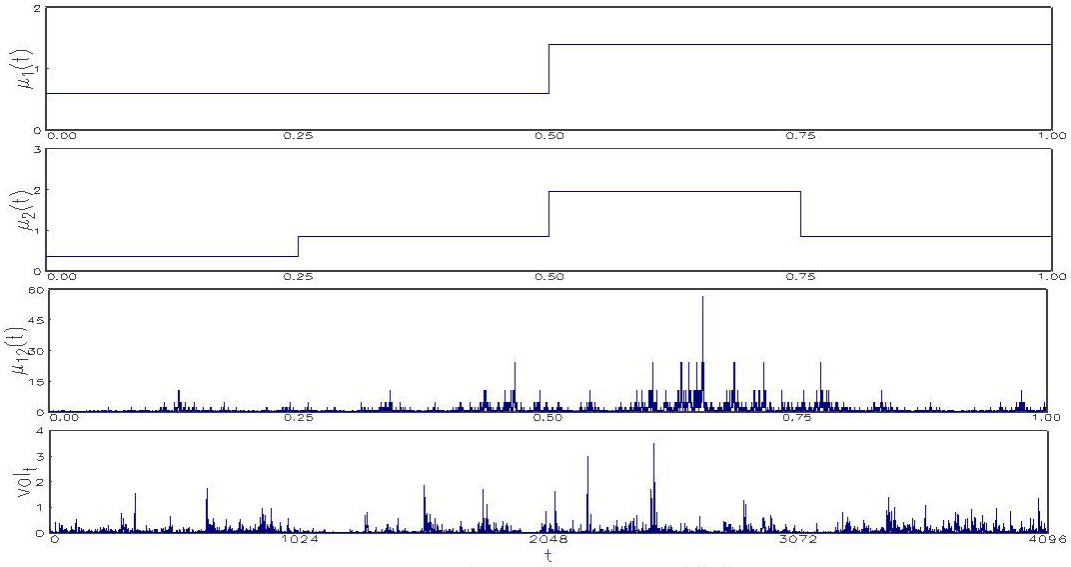

Figure 2 from page 52 of [1]

Many variations of this hierarchical procedure exist. One could think of generating $b \geq 2$ subintervals per iteration, for instance. Subintervals indexed from left to right by $\beta \in \{0, \dots, b-1\}$ receive mass $m_0, m_1, \dots, m_{b-1}$ and conserve mass by requiring $\sum_{\beta=0}^{b-1} m_\beta = 1$. Instead of always assigning the fraction m_0 to the left descendent, one could alternatively randomize this transformation by making a draw from a random variable M_β that takes values $m_0, m_1, \dots, m_{b-1}$ with probabilities $p_0, p_1, \dots, p_{b-1}$, or for that matter, from a more general random variable $M \geq 0$. A popular example of this generalization is the Lognormal Multifractal model, where M obeys a Lognormal distribution [36,47]. Note that these factors remain independent across the cascade steps. Mass conservation within a cascade step, on the other hand, is fulfilled by setting a constraint in the joint distribution of $\{M_0, M_1, \dots, M_{b-1}\}$ so that $\sum_{\beta=0}^{b-1} M_\beta = 1$.

A further extension in the construction conceives the independence of $\{M_0, M_1, \dots, M_{b-1}\}$ within each cascade step. M_β may be allowed in this case to have the same distribution law as the random variable $M \in \{m_0, (1-m_0)\}$, or again, more generally, of $M \geq 0$. Note that when the independence of $\{M_0, M_1, \dots, M_{b-1}\}$ is considered, and mass concentration can only be assured on average, that is, provided $M_\beta \stackrel{\text{def}}{=} M$: $\mathbb{E}\left[\sum_{\beta=0}^{b-1} M_\beta\right] = 1 \Leftrightarrow \mathbb{E}[M] = 1/b$. One refers in this case to a canonical measure.

At last, let us verify now how the last two constructions satisfy the definition of multifractality. We note that at each starting point $t = \sum_{i=1}^{k} \eta_i b^{-i}$ with $\eta_1, \eta_2, \dots, \eta_k \in \{0, \dots, b-1\}$, the conservative measure μ_k in the interval $\Delta t = b^{-k}$ is $\mu(\Delta t) = M_{\eta_1} * M_{\eta_1,\eta_2} * \dots * M_{\eta_1,\eta_2,\dots,\eta_k}$. The exact conservation of mass and the independence of multipliers across cascade steps lead furthermore to $\mathbb{E}[\mu(\Delta t)^q] = \mathbb{E}[M^q]^k$ or respectively to $\mathbb{E}[\mu(\Delta t)^q] = (\Delta t)^{\tau(q)+1}$, where $\tau(q) = -\log_b \mathbb{E}[M^q] - 1$, and where $k \to 1$ implies that $t \to 0$.

References

[1] Leövey A.E. (2015). *"Multifractal Models: Estimation, Forecasting and Option Pricing"*. Doctoral thesis Christian-Albrechts University, Kiel, Germany. Retrieved from http://macau.uni-kiel.de/receive/dissertation_diss_00016930

[2] Calvet L. E. & Fisher A. J. (2001). *"Forecasting multifractal volatility"*. Journal of Econometrics, 105, (1), 27-58.

[3] Calvet L. E. & Fisher A. J. (2004). *"How To Forecast Long-Run Volatility: Regime Switching And The Estimation Of Multifractal Processes"*. Journal of Financial Econometrics, v2(1, Winter), 49-83.

[4] Hainaut D. & Boucher J-P. (2014). *"Frequency and Severity Modelling Using Multifractal Processes: An Application to Tornado Occurrence in the USA and CAT Bonds"*. Environmental Modeling & Assessment, vol. 19, no 3, pp. 207-220.

[5] Boucher J-P. & Hainaut D. (2013). *"Time Series of Count Data using Multifractal Process"*. Preprint. Université du Québec à Montréal - UQAM, Montréal, Canada.

[6] López Cabrera B. (2010). *"Weather Risk Management: CAT bonds and weather derivatives"*. Humboldt University, Berlin, Germany.

[7] Public domain (2016). "*CCR compensation scheme for natural disasters"*. Caisse Centrale de Réassurance official website. France. Retrieved from https://www.ccr.fr/-/indemnisation-des-catastrophes-naturelles-en-france

[8] Lunney K. (2011). *"FEMA recoups millions in improper payments"*. a Government Executive article. Retrieved from http://www.govexec.com/oversight/2011/03/fema-recoups-millions-in-improper-payments/33561/

[9] Scott Arnold N. (2000). *"The Role of Government in Responding to Natural Catastrophes"*. Journal des Economistes Et des Etudes Humaines 10 (4) (2000). pg. 6.

[10] LeBlanc A. (2011). *"Managing the escalating risks of natural catastrophes in the United States"*. Online report from the "International Regulatory Affairs and Exposure Management" department at Lloyd's, U.S.

[11] AIR Worldwide source (2015). *"Gap Between Insured & Economic Losses Offers Business Opportunities"*. The Insurance Journal, Sept. 18.

[12] Munich Re source (2015). *"Rise in Weather Risks: When nature becomes a threat"*. Retrieved from Munich Reinsurance official website.

[13] Kinghorn J. (2008). *"Coastline at Risk: 2008 Update to Estimated Insured Values of US Coastal properties"*. AIR Worldwide report. Retrieved from AIR Worldwide official website.

[14] The Columbia Electronic Encyclopedia, 6th ed. (2012). *"Tornado definition and its features"*. Columbia University Press.

[15] Wikipedia The Free Encyclopedia (2016). *"Tornadogenesis"*.

[16] Wind Science and Engineering Center project (2006). *"Enhanced Fujita Scale"*. Texas Tech University, Lubbock, Texas, U.S.

[17] Weather reports (2000-2016). *"Fujita Tornado Intensity Scale"*. InfoPlease official website.

[18] Nathan Risk Suite (2016). *"Nathan World Map of Natural Hazards"*. Retrieved from Munich Reinsurance official website.

[19] Imbornoni A-M. (2000-2016). *"Tornadoes: Facts and figures about twisters"*. InfoPlease official website.

[20] DiCaprio L. (2016). *"Before the Flood"*. A National Geographic documentary on climate change issues. Watch at http://channel.nationalgeographic.com/before-the-flood/

[21] SHELDUS (2016). *"Tornado arrivals and damage costs"*. from the Spatial Hazard Events and Losses Database, South University of Carolina, SC, U.S.

[22] Public print (2016), *"STL decomposition of a Time Series"*. OTexts, online open-access textbooks. Retrieved from https://www.otexts.org/fpp/6/5

[23] Kedem B. & Fokianos K. (2002). *"Regression Models for Time Series Analysis"*. University of Maryland & University of Cyprus, Wiley, NY, U.S.

[24] Schulz J. (2013). *"Generalized Linear Models for a Dependent Aggregate Claims Model"*. Master's thesis for the Department of Mathematics and Statistics, Concordia University, Montreal, Canada.

[25] Barnett A.G., Baker P. & Dobson A.J. (2012). *"Analysing Seasonal Data"*. The R Journal Vol. 4/1.

[26] Geyer C. J. (2012). *"Introduction to Markov Chain Monte Carlo"*. Chapters 1 and 6. Retrieved from the online library
http://www.mcmchandbook.net/HandbookChapter1.pdf

[27] Wikipedia The Free Encyclopedia (2016). *"Continuous distributions"*.

[28] Wikipedia The Free Encyclopedia (2016). *"Skewness"*.

[29] Achieng O. M. (2011). *"Actuarial Modeling for Insurance Claim Severity in Motor Comprehensive Policy Using Industrial Statistical Distributions"*. Bosom Insurance Brokers LTD, Nairobi, Kenya.

[30] Engineering Statistics Handbook (2016). *"Quantile-Quantile Plot"*. Section 1.3.3.24 from NIST Sematech Online course.

[31] Charpentier A. (2016). *"Tarification et évaluation en IARD"*. a course given at Université du Québec à Montréal (UQAM).

[32] Public blog (2017). *"Link function vs inverse canonical link function"*. Retrieved from Stacke Exchange online official website.

[33] Riedi R. H. (2002). *"Multifractal Processes"*. Public print, Rice University Statistics, Houston, Texas, U.S.

[34] Mandelbrot B. B. (1963). *"The Variation of Certain Speculative Prices"*. The Journal of Business, 36(4): 394-419.

[35] Mandelbrot B. B. (1972). *"Possible Refinements of the Lognormal Hypothesis Concerning the Distribution of Energy Dissipation in Intermittent Turbulence"*. In M. Rosenblatt and C. Van Atta, editors, Statistical Models and Turbulence. Springer Verlag.

[36] Calvet L. E. & Fisher A. J. (2008). *"Multifractal Volatility"*. Academic Press, Amsterdam.

[37] Lecki, T. R., & Rutkowski, M. (2004). *"Credit risk: Modeling, valuation and hedging"*. Springer Finance.

[38] Hamilton, J. D. (1989). *"A new approach to the economic analysis of nonstationary time series and the business cycle"*. Econometrica, 57(2), 357–384.

[39] Kalman, R. E. (1960). *"A new approach to linear filtering and prediction problems"*. Journal of Basic Engineering, 82(1), 35–45.

[40] Baur P. & Breutel-O'Donoghue A. (2004). *"Understanding Reinsurance"*. A Swiss Reinsurance Company work, published by Economic Research & Consulting, Zurich, Switzerland.

[41] Boulliung F. (2017). *"Non-Life Reinsurance"*. from a Partner Reinsurance course given at the Strasbourg University.

[42] Munich Re (2010). *"A basic Guide to Facultative and Treaty Reinsurance"*. A Munich Reinsurance America, Inc. work, Princeton, NJ, U.S.

[43] McDonald R. L. (2006). *"Derivatives Markets"*. 2nd Edition Pearson, Northwestern University.

[44] Lopéz Cabrera B. (2010). *"Weather Derivatives: CAT bonds and Weather Derivatives"*. Public print dissertation, Humboldt University, Berlin, Germany.

[45] Tuo X. & Guo W. (2011). *"Flood insurance bonds pricing based on the Monte Carlo simulation"*. Hunan University of China, Systems Engineering Proceida (2) 199 - 204.

[46] Liu J., Xiao J., Yan L. & Wen F. (2014). *"Valuing Catastrophe Bonds Involving Credit Risks"*. Hindawi Publishing Corporation, Vol 2014, 563086, 6 pages.

[47] Mandelbrot B. B., Calvet L. E., &Fisher A. (1997). *"A Multifractal Model of Asset Returns"*. Discussion Paper 1164, Cowles Foundation.

[48] Confidential documentation of a Vendor model (2014). *"Severe convective storm modeling"*. Retrieved from the official training program for the Certified Catastrophe Risk Analyst.