Modeling of the subsidence risk in Metropolitan France

TOURRET Alice

ALLIANZ France

07/09/2020 – 31/08/2021

Acknowledgments

I would first like to thank each colleague of the P&C Reserving team for their warm welcome and their support during 15 months of internship and apprenticeship, through on-site and home working.

I especially want to thank Dominique Abgrall and Julia Simaku for their help and guidance on the project but also for their global perspective on the professional world.

Finally, I would like to thank Eve Leconte, my academic supervisor, for her support and follow-up throughout the academic year.

Table of content

Modeling of the subsidence risk in Metropolitan France

ABSTRACT

In the past decade, there has been an increase in the occurrence and severity of droughts throughout the French metropolitan territory. The subsidence risk, which is defined as land movements caused by droughts, has therefore become a growing topic of interest for P&C insurers. However, mechanisms behind droughts and subsidence claims are complex and depend on climatic, geotechnical but also governmental factors. For such reasons, using regular reserving methods to estimate the ultimate financial loss for the subsidence risk might lead to inadequate results. During this project, we propose two different models using GLM and XGBoost algorithms to predict the ultimate cost of subsidence for the years of occurrence 2016 to 2020. The predictions of our final model translate the severity of the different years of occurrence and provide a first early estimation for recent years of occurrence such as 2020. However, it is still difficult to predict accurately the ultimate loss even with external climatic and geotechnical data.

Introduction

Natural disasters are a growing concern for the insurance sector. Indeed, natural disasters are extreme events, meaning they are difficult to predict and costly to cover when they do occur. In August 2021, a report from the GIEC [1] warned that global warming is already responsible for the acceleration of the occurrence of extreme events, and that if there are no improvement in human carbon-related activities in the future, the frequency and the severity of these events will continue increasing. This means that, for insurance companies, the uncertainty around natural disasters is growing and more and more actions are being taken to understand, model and predict such risks.

Subsidence is the second most important natural disaster in France after floods in terms of loss according to the Caisse Nationale de Réasurance (CCR) [2]. The subsidence risk is a direct consequence of drought: land movement can occur as soil retract and extend due to a lack of precipitations. This can damage houses, entailing cracks in the wall or the ground but also sometimes leading to the weakening of the foundations of houses.

The compensation for subsidence claims is directly linked to the French natural disaster governmental process. Its goal is to check the abnormal nature of the event and to qualify it as natural disaster. However, such process can be long, and the claims often takes 1 to 4 years to be processed by the insurance company. The delay in the reporting process and its consequences on the available data lead to difficulties using traditional reserving methods, such as the Chain-Ladder method, commonly used by actuaries to model the ultimate financial loss of a year of occurrence. This can create uncertainties around the cost of severe years of subsidence such as 2018.

Using claim data from Allianz's database but also external climatic data, we propose and compare two different statistical models, using the algorithms GLM and XGBoost, to predict the severity of the subsidence of a year of occurrence right after the first year of development. Such a model can provide a first early estimate of the ultimate financial loss the company will be facing in the future.

In the first part, we define the subsidence risk as to understand how to model it. We begin with a presentation of the insurance of natural disasters in France. Then, we look thoroughly at the causes and consequences of the subsidence risk and its link to drought. To understand our database, we also study the context of the droughts between 2016 and 2020. Finally, I quickly explain traditional reversing methods and why they might not produce an accurate estimate of the cost of the subsidence risk.

In the second part, we model the ultimate loss caused by subsidence for each of the years of interest. In this part, we first have a look at the descriptive variables chosen and their data sources. Then, we define the context of the model and its characteristics. Finally, we compare the predictions and the robustness of two different algorithms, GLM

and XGBoost. We finish by computing the ultimate loss for each year of occurrence and interpreting the results.

Context of the project and presentation of the company

Allianz France is one of the main actors of the French Insurance market. I worked for around a year and a half in the Reserving Property and Casualty (P&C) Department of the Actuarial branch of the company. The department's role is to ensure that Allianz is reserving properly for all P&C (IARD) risks. Contracts in P&C cover properties (car, houses, equipment...) from risks of fire, thefts, car accidents among other. I had the opportunity to work on various topics, such as traffic accident mortality, non-affirmative cyber risk and finally the modeling of the subsidence risk. During the report, I will only talk about the latter. For confidentiality reasons, the numbers displayed in this report have been modified from the original results.

I. What is the Subsidence risk?

First, we will study the legislative and governmental context behind natural disaster insurance in France. Then, we will describe what is the subsidence risk and how it is related to drought. Afterwards, I will explain the context behind the different droughts of the years of occurrence 2016 to 2020, which are our years of interest. Finally, we will see why natural disasters and especially the subsidence risks are challenges for traditional reserving methods and why we may need to try different modeling methods to estimate the final loss of a year of occurrence.

A. Natural Disasters and the Insurance sector in France

France is one of the rare countries to have built a **legislative process** to guarantee a proper compensation in case of a non-insurable loss caused by a natural phenomenon. This compensation system was created in July 1982 with the law n°82-600 and relies both on Insurance companies and the government [3].

To be more precise, the Article L.125-1 of the French insurance code defines the effects of natural disasters as:

*"uninsurable direct material damage caused by the **abnormal intensity of a natural agent**, when the usual measures to be taken to prevent such damage could not prevent it from occurring or could not be taken".*

By natural phenomenon, here is a non-comprehensive list of the events that are considered:

- Floods
- Earthquakes
- Storms
- Mudflows
- Subsidence (including when caused by drought)

For this guarantee to apply, meaning for the victims to get the right for compensation, two conditions need to be filled. First, the victim must be subscribed to **a multi-risk housing insurance contract**. Within this contract, the "natural disasters" guarantee will be included and provide coverage for material damage to insured property such as residential buildings, professional buildings, furniture, cars, or equipment (crop and livestock included). Second, **an inter-ministerial decree** must have been published in the Official Journal. This decree declares a state of natural disaster within the municipality of the claim.

But how and why is a natural disaster state declared? It is important to understand how the natural disaster recognition system works, as it will directly impact the model we will build. Indeed, our data only contains claims that originated from municipalities that have been recognized in a natural disaster state by the French authorities.

In practice, here is the process: the mayor of a municipality that has suffered from a natural disaster submits a request to the prefectural services. The inter-ministerial commission, led by the Ministry of the Interior, is responsible for giving an opinion on the nature of the phenomenon and on whether its intensity is abnormal. This will be decided according to geotechnical and climatic reports that are provided with the request. The advisory opinion issued by the commission is then submitted to the government. When this decree is published, the victims have 10 days to contact their insurers and hand them the claim.

As expected for such a process, the delays can be quite long. It usually takes at least a year for a decree to be published. This is an issue for the insurance sector as it gives little to no visibility to the cost of a natural disaster event. We will talk more about this issue later.

B. Definition of the subsidence risk

a. The subsidence risk

As described in part I.A, subsidence is a natural disaster and is covered by the natural disaster guarantee on multi-risk housing insurance contracts in France. What are the factors and the consequences of subsidence?

Subsidence, or geotechnical drought, is defined as the **displacement of the ground surface** due to withdrawal-shrinkage of the soil [4]. This phenomenon is mainly caused by the clay composition of the soil and climatic conditions. Clay soil tends to shrink in dry weather and swells in humid conditions. Hence, severe droughts can lead to ground movements. This can be the cause of cracks in building walls or even a threat the integrity of the foundations of a building.

A lot of factors come into play concerning the occurrence of subsidence and its severity. When declaring a state of natural disaster, the Natural Disaster (CatNat) committee considers two criteria:

- **The geotechnical criterion**: soils that contain a high level of clay are the ones that are affected by subsidence. France is a country with a high degree of clay in the soil composition. However, we will see in part II.B.d that some regions of France are more likely to be affected by subsidence due to a higher level of clay in the soil.
- **The climatic criterion:** Subsidence mostly happens after a prolonged period of drought. We will study later different indicators that are used to measure the climatic and humidity conditions.

The type of housing is also an important factor to subsidence. It is important to point out that almost only **individual houses** are affected by subsidence [4]. Indeed, houses tend to have lighter and more flexible foundations than higher buildings. This can make them more vulnerable to land movements and subsidence.

Subsidence is the second most costly peril of the CCR catastrophe regime after floods [2]. However, the risk is considered to be hard to model. One of the main issues is that subsidence is a risk with usually a long declaration period due to the long CatNat process that declares or not the validity of the claims. For subsidence claims, **the average report period is 18 months** while it is 50 days on average for other natural disasters [5]. Another issue is that droughts from year to year are very different in term of intensity, duration and month of occurrence, which leads to difficulties predicting atypical and prolonged droughts such as the one of 2018.

b. Subsidence: a consequence of drought

The most important factor to the subsidence risk is unusually dry weather conditions alias drought. To model the subsidence risk, it is important to consider what is drought and how it is characterized.

Drought is defined as a **prolonged period of abnormal water shortages**, whether atmospheric (lack or precipitation), surface water (for agriculture purposes) or ground water [6]. Drought can be considered as a natural disaster, as it has a low occurrence frequency and a high severity. However, at the contrary of storms, floods or other disasters, the damages caused by the drought are often scattered over a several months.

Technically, drought is mostly characterized by:

- A **low amount of precipitations** over a certain period
- **High temperatures** during a certain period

Several **indexes** have been created and used to measure drought over the past decades. We will quickly go over the ones that will be useful in the case of our analysis: the Standardized Precipitation Index (SPI), the Standardized Precipitation and Evapotranspiration Index (SPEI) and the Soil Water Index (SWI). Let us have a first look at the indexes [7]:

- **SPI**: The most widely used drought index, as the only needed information to compute it is precipitation data. The SPI quantifies the deviation of the precipitations of the period from the historical average. Computing the SPI requires to calibrate a Gamma law on 30 years of historical precipitation data or more. The R package *SPEI* can help speed up the computing process. A SPI of 0 indicates that the period is normal in terms of precipitations. A positive SPI indicates a rather dry weather while a negative SPI a humid weather. The more the SPI strays away from 0, the more the intensity of the event is.
- **SPEI**: The SPEI is an extension of the SPI, using precipitation but also temperature data as to evaluate evapotranspiration. Evapotranspiration is the quantity of water that evaporates from the soil. Drought is often characterized by a high evapotranspiration, as high temperatures lead to dry soil. The computing and interpretation process is similar to the SPI index.
- **SWI**: The SWI indicates the humidity level of the soil [8]. It is quite different from the SPI and SPEI as they measure atmospheric drought rather than geotechnical drought. The index varies from 0 to 1, 1 being an extremely wet soil and 0 an extremely dry soil. The monthly SWI is computed by Météo-France as a part of the CatNat process. We will talk more about it in part II.B.e.

For the insurance industry, **climate change** is a rising issue as it increases extreme events (droughts, floods...) in terms of frequency but also severity. Research have already been done on the connection between drought and climate change. Iglesias et al. [9] exposed that climate change could increase the risk of drought in the future.

This partly explains the increasing interest of insurance companies in modeling natural disasters including subsidence claims. In 2021, the Mission des Risques Naturels [10] pointed out that almost **40% of the cost** of subsidence claims since 1989 are concentrated over the **period 2015-2019**. In part I.D, we will also observe this acceleration in Allianz's claim data. However, the effects of climate change on the frequency and the severity of natural disasters are be hard to model and quantify as there are a lot of uncertainties around the phenomenon.

C. Droughts in France: 2016-2020

During this project, we will study the years of occurrence 2016 to 2020. It is important to know the context of drought behind each of these years of occurrence if we want to model properly and understand our results.

Indeed, one of the difficulties behind modeling drought is that droughts have different characteristics from one year of occurrence to another. The differences can be:

- **On a geographical scale:** depending on the year of occurrence, different parts of France can be affected by drought. For the subsidence risk, there is a high geographical dependency: claims are not distributed equally over the French territory. We will see that for each year of occurrence, different departments are affected by drought and subsidence.
- **On a temporal scale:** Depending on the precipitation and the temperature patterns, drought and subsidence can happen in different months and different seasons.

The years of occurrence 2016 to 2020 are quite **particularly severe years for drought**. As said before, the Mission des Risques Naturels [10] pointed out that almost 40% of the cost of subsidence claims since 1989 are concentrated over the period 2015-2019. Each of the five years of occurrence are considered as severe years for the subsidence risk. However, some are more severe than others such as **2018**, which ultimate cost is still unknown according to the CCR due to uncertainties linked to the severity of the events [11].

First, let us have a look at the distribution of the municipalities declared in a state of natural disaster over the French territory by year of occurrence. All the following

information and data have been retrieved from CCR reports [11] [12] [13] [14] [15], which is a trustable source for natural disaster events usually used by insurance companies.

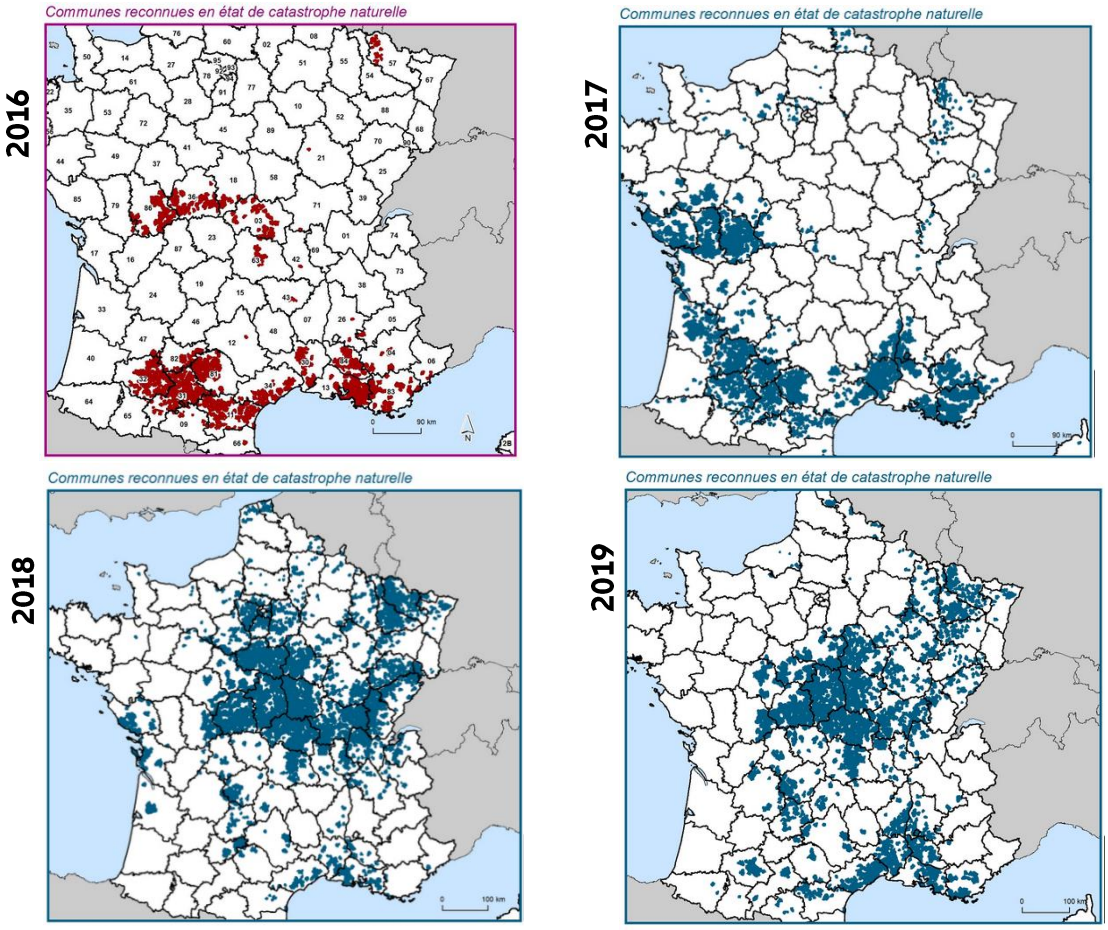

Figure 1: Distribution of the municipalities declared in a state of natural disaster over the French territory by year of occurrence (2016-2020)

Year of occurrence	2016	2017	2018	2019
Number of municipalities	986	2 122	4 060	2 905

Table 1: Number of French municipalities declared in a state of natural disaster by year of occurrence (2016-2020)

Let us have a look at each year of occurrence individually. First, 2016 is not as severe in term of drought relatively to the other years of occurrence considered. The number of municipalities considered as in a natural disaster state in quite low (below 1000). The affected municipalities seem to spread in the south from the Haute-Garonne

department and its surroundings to Marseilles (Occitanie, Provence-Alpes-Côte-d'Azur). The center of the country and the east (Massif Central) are also affected. 2016 recorded especially dry conditions in July, August, September and December.

2017 follows the same pattern as 2016: southern and center-east departments are affected by drought. However, the severity of the drought seems to be higher. There are two times more municipalities declared in a natural disaster state than in 2016! We can expect to observe more claims from the occurrence 2017 than 2016, but these claims will be distributed on the French territory in a similar manner. The drought of 2016 first started in April, before surplus precipitations in June. However, it resumed for the summer, with exceptionally dry conditions in August.

2018 is renowned in the P&C insurance sector as an **"atypical year"**. When looking at the map of municipalities in a natural disaster state, the municipalities of interest are distributed quite differently than in 2016-2017. In 2018, the most affected areas seem to be the center of France, the Parisian region and center-east/north-east. The severity also seems a lot higher: 4 060 municipalities are considered in a natural disaster state, more than 4 times the number of municipalities than in 2016. To this day, the CCR warns that the CatNat process is still not over even though we are on the third year of development. This means that the number of municipalities that are considered affected by drought in 2018 might still increase: there is still a significant uncertainty on this year of occurrence due to its severity. Concerning climatic circumstances, 2018 had an extremely dry summer for the overall French Metropolitan territory such as in 2016 and 2017. However, contrarily to the past years of occurrence, the drought lasted until December for some areas in France, making it an extremely prolonged event. in November 2018, Provence-Alpes-Cotes-d'Azur and Languedoc-Roussillon recorded surplus precipitations, more than twice the average monthly precipitations. This can explain why even though 2018 is a record year for drought, the regions that are typically affected by the event were not affected during this year. This reinforced 2018 as an atypical year that might be hard to model.

2019 follows a similar geographical pattern as during 2018, with also the center-south of France being affected. The severity seems to be lower. However, it is important to keep in mind that we have only two years of development for 2019, and more municipalities might be recognized in a natural disaster state soon. Dry conditions for 2019 started in January and lasted until September for the disaster-stricken areas.

Concerning 2020, one year of development is observed and the average CatNat process takes more than two years, meaning the current map is only temporary and is not

displayed here. It seems however that 2020 follows the same pattern as 2018 with a lowered severity [15].

It is possible that these differences between each year of occurrence will make it difficult to model the subsidence risk with only five years of historical data.

D. Chain ladder: traditional reserving methods and extreme events

The goal of P&C Reserving departments in Insurance companies is to **estimate the ultimate cost of the different P&C risks** (natural disasters, car accidents, fires...) the company is facing. The estimations are used to reserve, meaning putting aside an appropriate amount of financial funds, to properly face these risks. As we have seen before with the CatNat process, the claims and their cost are often not reported to the insurance company within the year of occurrence, but often one to several years afterward. Hence one of the goals of reserving is being able to anticipate these upcoming claims and their cost. Through reserving, we estimate an ultimate loss for a year of occurrence, meaning the final cost including the cost of the claims that are still unobserved.

Before starting, let us take a factice simple example to understand the Reserving vocabulary and methods. For the risk of subsidence and the year of occurrence 2021, we observe 20 claims which evaluate to 500 000 euros. However, 2021 is only what we call the first "year of development" of the occurrence 2021. As we have seen before, subsidence claims can take more than two years to be reported to the insurance company due to the CatNat process. This means that most claims of the occurrence 2021 will be reported in year of development 2, which is 2022, or year of development 3, which is 2023. The question at the end of 2021 is: how much is left to pay for this occurrence? This amount is called IBNR ("Incurred But Not Reported"). The sum of the observed loss and the IBNR is called the ultimate loss, or simply the "ultimate". Let us say that the department estimates the IBNR for the subsidence risk of occurrence 2021 to be 4 500 000 euros, the estimated ultimate loss will be five million. But how do we estimate the incurred but not reported loss and the ultimate loss?

One of the most common models to estimate the ultimate cost of claims is the **Chain-ladder method**. We first quickly explain the theory of Chain-ladder and the assumptions the model relies on.

The Chain-Ladder method is a projection of the losses in the future computed from the losses observed in the past [16]. To run this method, we first need to aggregate the claim data by year of occurrence and year of development, the year of development being the year when the claim was reported to the insurer. We use the aggregate to create a loss triangle. To be clearer, let's have a look at a theoretical triangle. We write $L_{i,j}$ the total

loss observed at year j for the claims occurred at year i. We say that there are n years of occurrence.

	Year of development					
Year of occurrence	1	2	…	j	…	n
1	$L_{1,1}$	$L_{1,2}$	…	$L_{1,j}$	…	$L_{1,n}$
2	$L_{2,1}$	$L_{2,2}$	…	$L_{2,j}$	…	$L_{2,n}$
…	…	…	…	…	…	…
i	$L_{i,1}$	$L_{i,2}$	…	$L_{i,j}$	…	$L_{i,n}$
…	…	…	…	…	…	…
n	$L_{n,1}$	$L_{n,2}$	…	$L_{n,j}$	…	$L_{n,n}$

Table 2: triangle of cumulated loss by year of occurrence and year of development

For each year of occurrence i between 1 and n, the goal is to compute the ultimate loss $L_{i,n}$. To do so, we can use the following formula:

$$L_{i,n} = L_{i,1} * F_{i,1} * \ldots * F_{i,n-1} \text{ with } F_{i,j} = \frac{L_{i,j+1}}{L_{i,j}} \qquad (L_{i,1} \neq 0)$$

However, for the lower left triangle (cells that are filled with red in the triangle below), the loss is not observed, meaning we cannot directly use this formula.

	Year of development					
Year of occurrence	1	2	…	j	…	n
1						
2						
…						
i						
…						
n						

Table 3: Simplification of the loss triangle

Before using the Chain-ladder method, we need to make several assumptions:

1. The years of occurrence are independent.
2. There exist parameters $f_1, \ldots, f_{n-1}$, called development factors, such that $E(F_{i,j} | L_{i,1}, \ldots, L_{i,j}) = f_j$ for $1 \leq i \leq n$ and $1 \leq j \leq n-1$.

3. There exist parameters $\propto \in \{0,1,2\}$ and $\sigma_1, \dots, \sigma_{n-1} > 0$ such that $var\left(F_{i,j} \middle| L_{i,1}, \dots, L_{i,j}\right) = \frac{\sigma_j^2}{L_{i,j}^2}$ for $1 \leq i \leq n$ and $1 \leq j \leq n-1$.

Using the assumptions previously made, the development factors can be computed by:

$$\hat{f}_j = \sum_{i=1}^{n-j} \frac{L_{i,j}^{\propto}}{\sum_{k=1}^{n-j} L_{k,j}^{\propto}} F_{i,j}$$

Usually $\propto$ is set to one. We can then use the development factors in the first ultimate loss equation for each year of development $j > n - i + 1$.

To illustrate this method, we test it directly on our data. To be able to estimate the ultimate loss, we need more than five years of historical data, so we take the claim data that is available, on the period 2001-2020. The following figure presents the IBNR loss by year of occurrence and year of development.

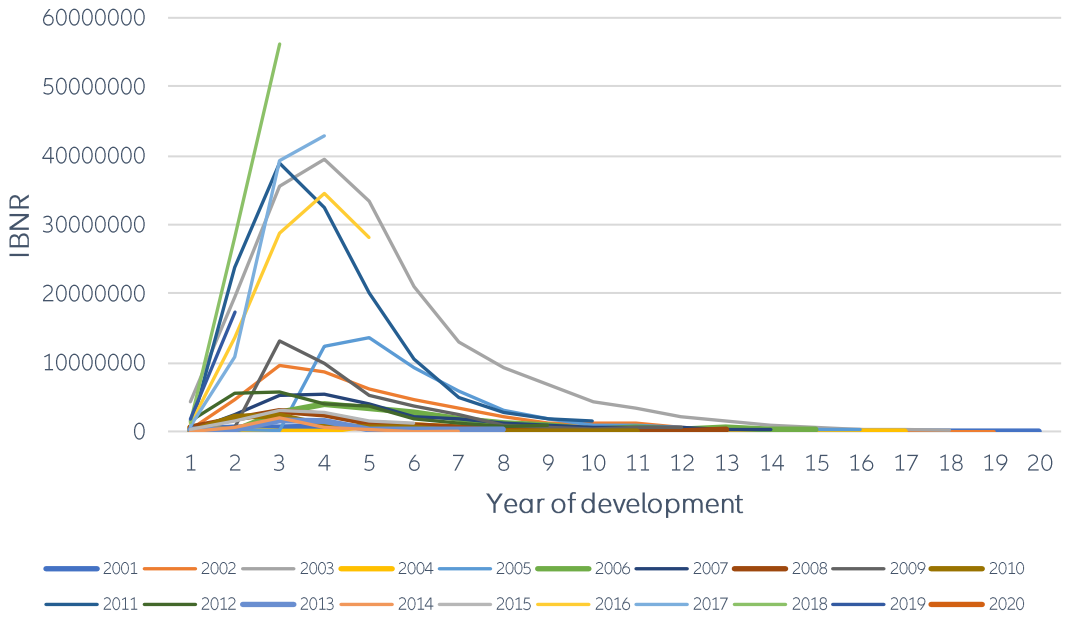

Figure 2: IBNR by year of occurrence and year of development

One can see that for the first years of occurrence such as 2001-2010, the IBNR are close to 0 for a high number of years of development, meaning we know the ultimate claim cost of these years of occurrence (all claims have been reported). We call these years of occurrence as **"mature".** They are important to be able to run the Chain-ladder method properly. At the opposite, the recent years such as 2017 and 2016 still have very

high IBNR in later years of development. The goal of the Chain-ladder method is to use the data available in the "mature" years of occurrence to be able to estimate the ultimate loss of the years which are not yet mature.

The next figure is the same data but represented in a triangle form by number of claims by year of occurrence and year of development. The data has been transformed for confidentiality reasons.

Year of development

Year of occurence	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2001	140	65	9	6	7	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0
2002	114	688	766	159	26	3	28	2	3	0	1	1	0	0	0	0	0	0	0	
2003	1052	2857	3618	1986	216	74	11	5	8	1	2	0	0	0	1	0	0	1		
2004	404	48	10	3	43	9	6	0	0	0	0	0	0	0	0	0	0			
2005	226	68	10	1403	43	17	9	3	1	1	1	1	0	0	0	3				
2006	85	9	332	26	3	3	5	1	0	0	0	0	0	0	0					
2007	31	399	85	20	14	1	1	0	1	0	0	1	0	0						
2008	218	136	55	12	4	0	0	0	1	0	0	0	0							
2009	89	21	743	24	2	3	0	0	0	2	0	0								
2010	37	115	43	3	5	0	0	0	0	0	2									
2011	99	3315	546	50	11	5	3	2	2	0										
2012	345	726	114	12	8	3	3	2	0											
2013	78	37	7	2	0	2	0	0												
2014	21	39	23	5	0	0	0													
2015	58	270	166	23	4	0														
2016	115	2063	350	30	7															
2017	260	2981	496	55																
2018	480	5232	371																	
2019	664	2796																		
2020	535																			

Figure 3: Triangle of non-cumulated reported claims by year of occurrence and year of development

Looking at the non-cumulated number of claims triangle, we can **observe unusual shocks** between years of development. For year 2005, 2006 and 2009, there are sudden significant recordings of claims at year of development 4 and 5. This can be considered as abnormal as these years seemed to have stabilized already. For example, for 2005, in year of development 3, only 10 claims were recorded. But for year of development 4, 1403 claims were recorded!

The reason behind this issue is the prolonged CatNat process, as described before. It is possible that the shocks correspond to the **publication of one or several Natural disaster decrees** that declare several municipalities in a disaster state. This decree entails that Allianz will receive numerous claims at the same time, hence **a group effect** can be observed.

The issue with these observed shocks between years of development is that it will obstruct with Chain Ladder's estimations. Indeed, the development factors can still be

computed but their results might not be satisfactory, as the lines of the triangle are quite different for each year of occurrence. Chain ladder's assumptions imply that the time period between claims should be independent. Here the Cat Nat process definitively makes it difficult to respect such assumption.

Even though we know the results might not be optimal, let us make a first estimation of the ultimate loss and ultimate number of claims using the Chain Ladder Method. The results can be found in the table below.

Year of occurrence	Net Loss Estimate
2001	1 662 413
2002	22 103 442
2003	96 707 020
2004	1 929 703
2005	27 165 897
2006	7 673 738
2007	13 380 036
2008	5 221 544
2009	24 539 922
2010	4 683 847
2011	105 701 140
2012	18 092 287
2013	3 571 618
2014	2 930 335
2015	13 680 728
2016	122 412 390
2017	143 295 268
2018	221 641 835
2019	158 720 172
2020	109 142 413

Table 4: Ultimate loss – Chain Ladder estimation

For 2018, we know that the technical direction is estimating the ultimate loss to around 285 million euros. Here, the method seems to underestimate the ultimate loss. It is possible that Chain Ladder provides an optimist prevision.

Chain-ladder is not the only traditional reserving method we could use to estimate the ultimate loss of the subsidence risk: occurrence frequency, Bornhuetter-Ferguson method... However, the literature shows that these methods often fail to reflect the extraordinary aspect of natural disasters [6] [17]. To estimate a more accurate ultimate loss, we will now try different methods that are used in extreme event modelling.

II. Modelling of the subsidence risk in metropolitan France

A. Goal and context of the model

As described in part I, insurers have little visibility of the subsidence risk. Indeed, due to the long CatNat process, insurance companies receive the subsidence claims on average 18 months after the event [5]. This makes it difficult to estimate the ultimate loss of a year of occurrence.

Furthermore, the CatNat system induces **a geospatial dependance** between claims. Indeed, when a decree declares a municipality in a natural disaster state, numerous claims can be submitted to the insurers at once which can result in a "group" effect. This is troublesome as it disqualifies the traditional reserving methods. It also means that it is difficult to estimate the ultimate loss from the observed reported claims of the first year of development, as they might not be linked to future natural disaster decrees.

However, we do know the **yearly climatic conditions and the severity of the drought** on the French territory by the end of the **first year of development** through data released by the National French Weather Agency, Météo France. Hence, we can use external climatic data to compute **a first estimate of the ultimate loss** of the subsidence risk.

As presented before in part I.D, the number of claims seems to stabilize at maximum on the year of development four for each year of occurrence considered. Hence, our model will focus on predicting **the number of claims for a year of occurrence at development four** with external climatic data available at the end of the first year of development. We will model at the **department** level. We will then add a **tail coefficient** to estimate the ultimate number of claims. To estimate the ultimate loss for each year of occurrence, we will multiply the number of claims of each occurrence year by a value range of **average costs of claims**.

In this part II, we will first have a look at our different data sources. We will rely on internal claim and portfolio data, but also, on external climatic but also geotechnical data. After that, we will detail our model and explicative variables. Finally, we will compare two algorithms, their predictions and estimate the ultimate loss for each year of occurrence.

B. Data Sources and Data aggregation

For this project, a lot of different data sources were used. In this part, we will describe them thoroughly with the process used to retrieve the variables of interest at the department level. We used in total 5 different data sources:

- **Allianz's MRH insurance contract portfolio** was used to understand the exposure to drought and subsidence.
- **Allianz's claim data**: the historical claim database which record claims and their losses starting from year of occurrence 2001 to 2020.
- **External climatic data** from the **ECA&D** to retrieve daily precipitations and temperature.
- **External climatic data** from **Météo France** to retrieve the SWI, the climatic index used in the Cat Nat process.
- **External soil composition data** from the French Bureau of Geological and Mining Research. This data is also used by the CatNat process (geotechnical criteria).

a. Exposure-to-risk: Portfolio data

First, we need to understand Allianz France's exposure to the risk of subsidence. To do so, we retrieved **the MRH portfolio data**. In this portfolio, every observation is a housing insurance contract. As said before, damage caused by subsidence is insured when the municipality where the claim is located is considered in a natural disaster state by the national French CaTNat committee. This means that **every contract in the portfolio may be affected by the subsidence risk**. Hence, measuring our exposure can be simply done by counting the number of contracts in Allianz's portfolio.

There are five databases available, one for each accounting year (2016 to 2020). For confidentiality reason, I will not directly display the numbers, but I will comment on the results of the introductory analysis.

The portfolio is quite stable from year to year between 2016 and 2020 with a slight decrease in terms of number of contracts. We want to look at the distribution of the contracts over the French metropolitan territory to understand which areas are the most prone to the subsidence risk (meaning where Allianz may see a lot of subsidence claims). This is important to have a look at because we will try to model the number of claims by department. However, this variable does not only depend on climatic conditions, but is **by construction limited by the number of contracts** of Allianz in the department. For example, a region may be deeply affected by drought and subsidence during a year, but

if it is not a region with a significant exposure, there will not be a lot of claim in the area. Hence, we need to include a variable such as the exposure within the model.

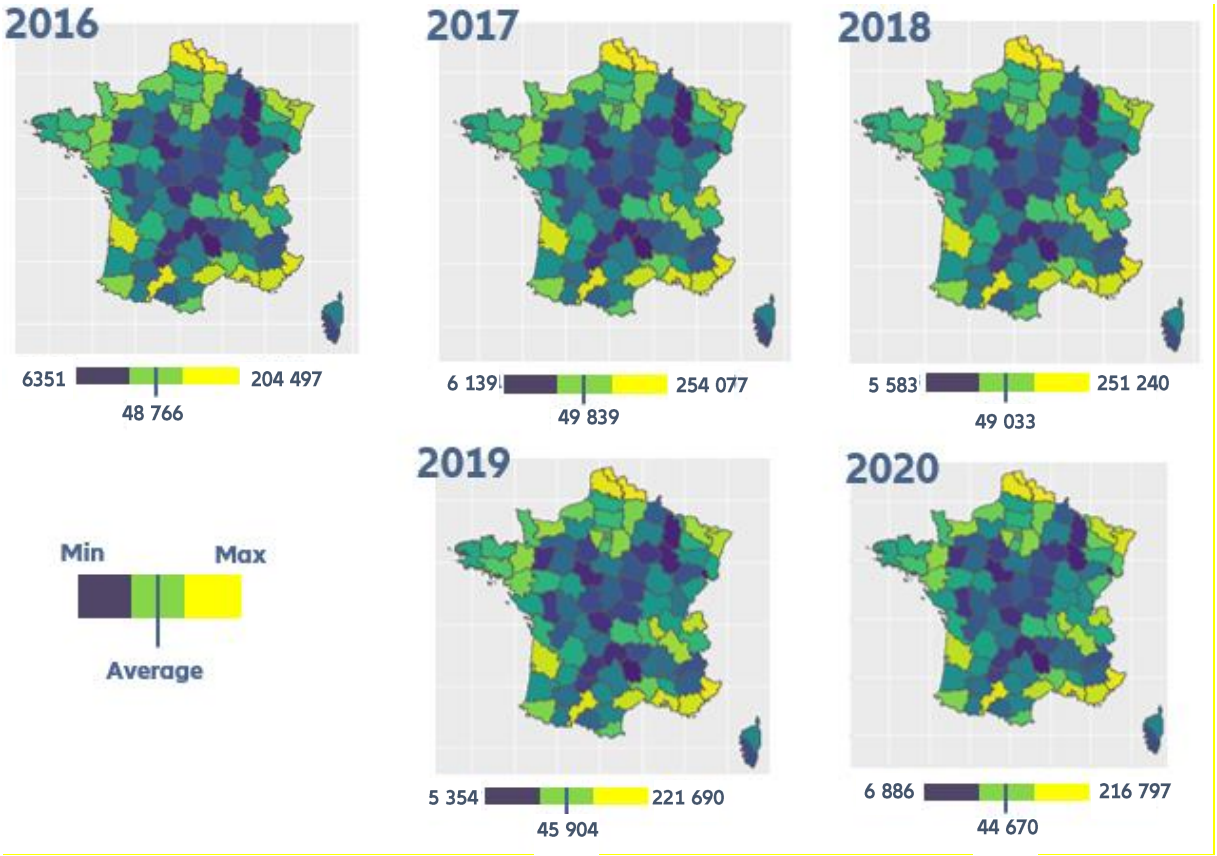

Figure 4: Map of the exposure by year of occurrence (2016-2020)

Looking at the above maps, we can confirm that the portfolio is stable year to year There is no movement between departments or huge change in numbers between 2016 and 2020.The departments that are the most exposed are in the North (Lille, Belgian borders), the Mediterranean coast, the big cities (Bordeaux, Lyon, Toulouse, Paris) and at the German and Swiss borders.

From this database, we retrieve the **exposure by department**. We also retrieve the **ratio of houses** within the department. As described in part I.B, houses are more affected by subsidence than higher buildings as the foundations are not as sturdy. Let us have a look at the distribution of houses over Allianz's portfolio:

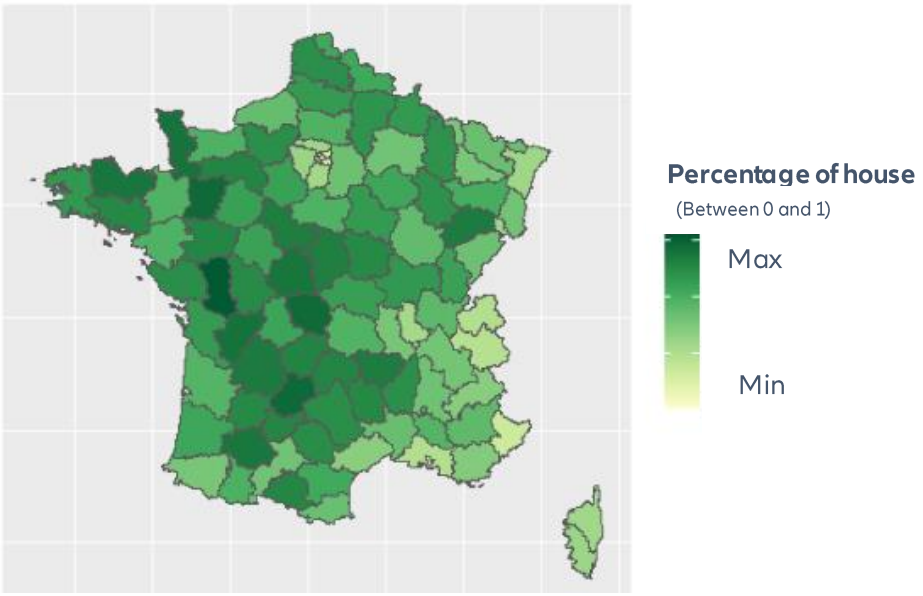

Figure 5: Map of the house percentage of housing insurance contracts of Allianz

Departments with the highest percentages of houses within the contract portfolio tend to be quite **rural departments**, certainly due to a low urbanization. They are mostly located in Bretagne but also the South-West. This is especially important to take note, as Occitanie is often affected by drought as seen in part I.C. At the contrary, it is likely than departments with a low number of houses (such as the Parisian departments) will not be highly affected by subsidence.

b. Claims data

To predict the number and the cost of subsidence claims by department, it is essential to retrieve company data on claims. We chose the years 2016-2020 as it is the years where we have portfolio data available, but we do have data on subsidence claims since the year of occurrence 2001.

Between 2016 to 2020, we observe **19 722 subsidence claims**. In the table below, one can find the distribution of claims by year of occurrence and insurance policy:

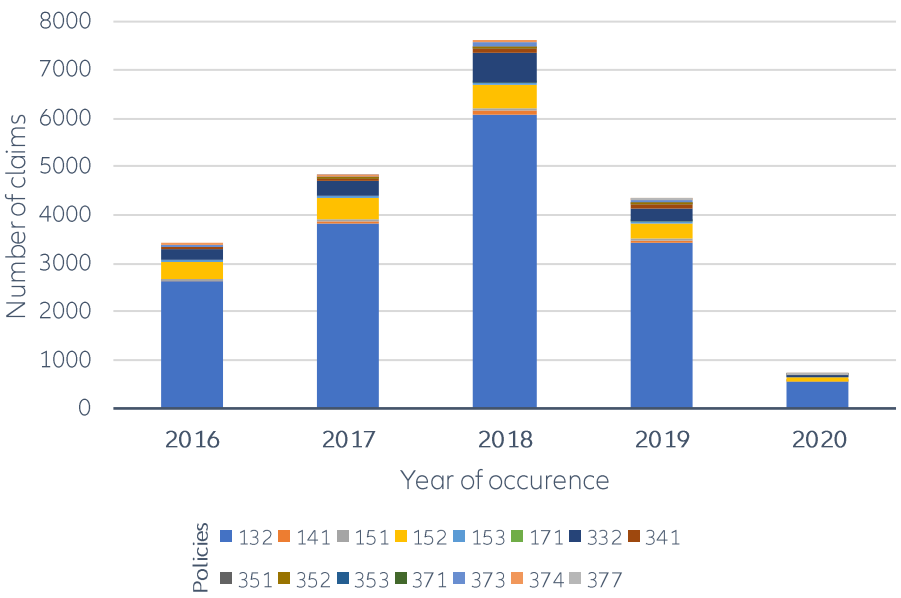

Figure 5: Number of claims by line of business and year of occurrence

Around 80% of all the claims are from the line of business **Multi-risk home insurance Agency** (132: MRH: assurance multirisque habitation Agence). This makes sense as the subsidence risk usually impacts constructions as seen in part I.B.a. The line of business Multi-risk home insurance Brokage (332) is also rather impacted. To simplify our analysis, we decided to only keep claims from the line of business 132, multi-risk home insurance for Agency. This filter limits our sample to 16 435 observations, ie subsidence claims. Looking at figure 6, we can study the number of our claims of interest by year of occurrence:

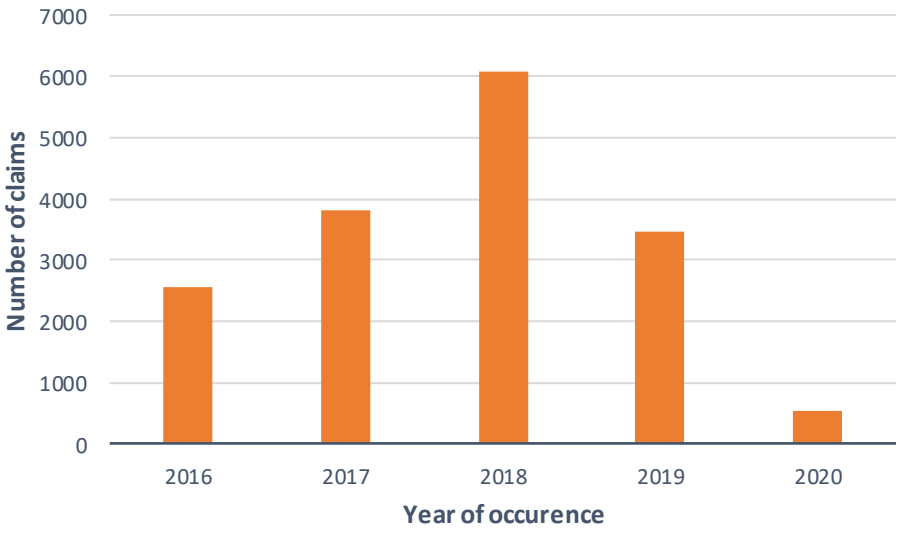

Figure 6: Number of claims of the segment 132 by year of occurrence

The year **2018** is by far the most affected by subsidence with more than 2 876 claims. This seems rather logical as we have seen in part I.C that it was a year with a severe and atypical drought in France. Furthermore, at the 31/12/2020, we are only on development year 3 for the year of occurrence 2018, so we expect this number to grow by the end of 2021. The same can be say for the years of occurrence 2019 and 2020, which are respectively on year of development 2 and 1.

Grouping our data at the department level, we can check if Allianz subsidence claim data is similar to the CCR climatic reports on drought for the years 2016-2020:

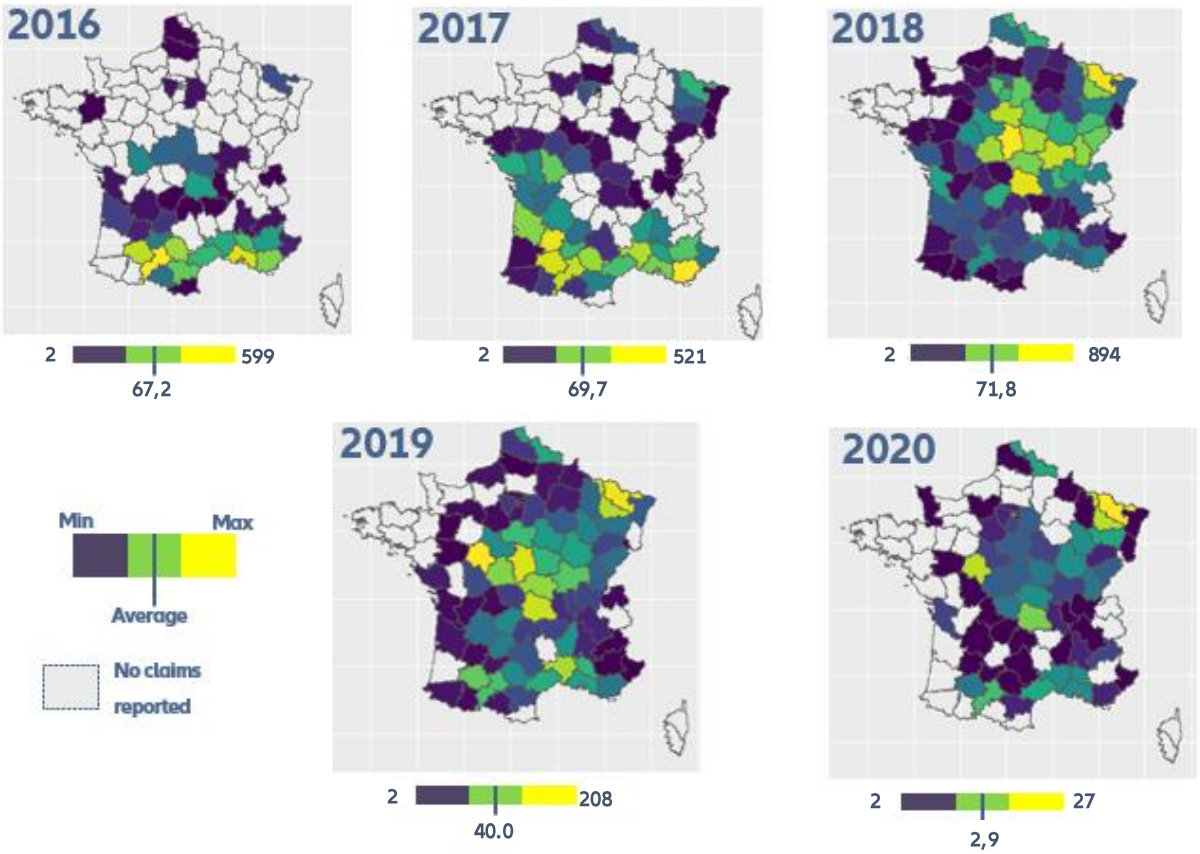

Figure 7: Map of the number of claims by department and by year of occurrence

2016 and 2017 are characterized by a high number of claims in the **South** while 2018-2020 in the **Center/North-East**. This analysis checks out with the CCR reports seen in part I.C. The few differences, for example a lack of claims in the center-west, might come from Allianz's own exposition, as the departments where there are few contracts do not record of a lot of claims.

c. Climatic data

Through Météo France, weather data is usually available quite fast after a year ends. Using already available climatic data, we want to have an approximation of the degree of gravity of the drought of the year to determine the number of claims Allianz might be facing early on.

In part I.B.b, we have already seen that drought is mostly characterized by:

- The average temperature
- The amounts of precipitations

To retrieve the climatic data, we use **the E-OBS grid dataset**. It is a public dataset published by the European Climate Assessment & Data project (ECA&D) [18] with the support of the European Commission. It is a time series dataset on a daily scale based off a 0.25*0.25 grid of the European territory. Zooming on France, we get a grid made from 61 * 41 = 2 501 cells, that we will call "stations". For each station, we have the following values for each day between the 01/01/1960 and the 31/12/2020:

- Average temperature
- Minimum temperature
- Maximum temperature
- Sum of precipitation

The following maps are an example of monthly-aggregated data:

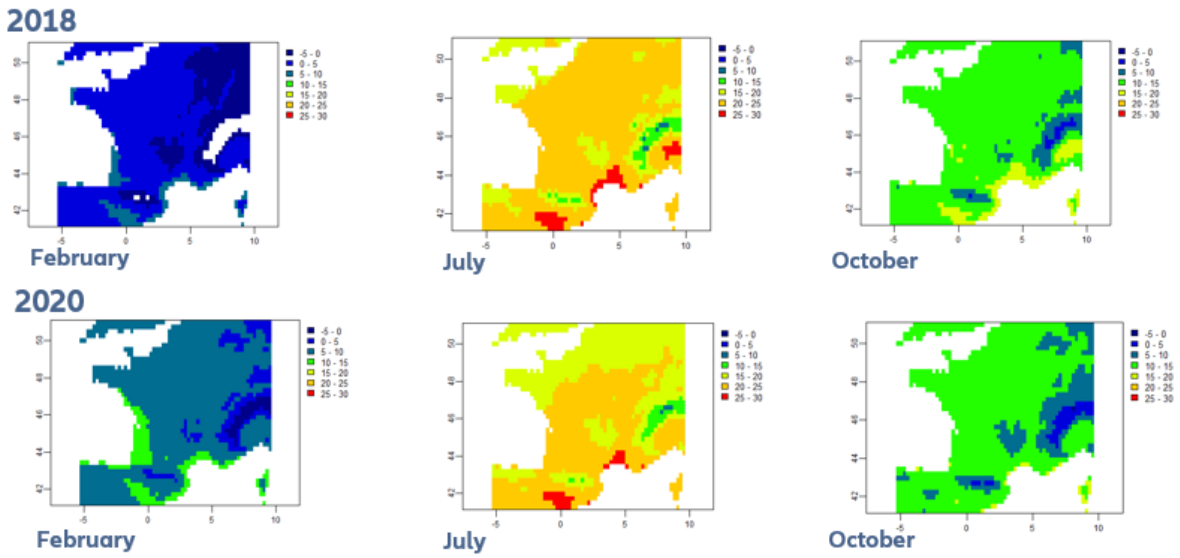

Figure 8: Several examples of monthly average temperature in France, 0.25*0.25 grid

The main issue with this dataset is its size and its structure. We want to simplify this format and to be able to translate this into usable information for our model and

prediction. As the cleanest scale of our portfolio and claim data is at the department level, we decided on aggregating the grid data at the department level, rather than the city of the original grid level. We keep the daily data, but we also compute monthly and seasonal aggregates that we will test during the feature selection.

Aggregation at the departement level

As a side note, aggregating the grid data to a department level is not a trivial task. We decided on linking the coordinates of the center of each cell of the grid to the nearest municipalities from the INSEE city code dataset. Then, we computed the composition in terms of departments (linked to municipalities) for each cell. For example, cell 130 is composed of 33% of cities in department 1, 33% cities in department 2, and 33% cities in department 3. We then computed the weighted average daily temperature and weighted average daily sum of precipitations for each department. For minimum and maximum temperatures and sum of precipitations, we simply take the minimum/maximum of the cells located within or at the boundary of the department.

Climatic indexes

We compute climatic indexes department-aggregated data. We compute the monthly and seasonal SPI and the SPEI indexes, described in part I.B.b. They need to be calibrated on more than 30 years of historical data. Hence, this aggregation was made on the time span 1980-2020, resulting in significant computational time. We chose to calibrate on 40 years to have a more time-efficient program.

d. Soil composition data

As described in part I, soil composition and especially **the presence of clay in the soil** is an important factor of subsidence. Hence, we need to include geotechnical specifications in our model. The French Bureau of Geological and Mining Research (Bureau de Recherches Géologiques et Minières, BRGM) published in 2016 a **mapping of the subsidence risk in France based on the clay composition of the soil** [19]. The level of hazard (1 - low, 2 - medium or 3 - high) is precisely evaluated on a grid of each French metropolitan department. For each department, we computed the mean hazard level and then created a table at the national level

According to our data, the south-east and south-west is particularly exposed to the subsidence risk due to soil composition. The northern borders with Belgium and Germany/Luxembourg and the upper center of the country are also quite exposed.

On the technical side, the data was only available at the department level on a shapefile (SIG) format, meaning there was some work to do to reconstitute the data into a national level table.

e. Soil Wetness Index

The soil Wetness Index is well documented in the literature [8]. Furthermore, it is used by Météo-France to communicate with the interministerial disaster commission as part of the Institution's contribution to the Natural Disaster system [8]. This means that this index plays a huge role in whether a municipality is declared in a natural disaster state after being hit by a drought, alias if the damages to houses in this area can be covered by the natural disaster statement of the insurance policy. But what is exactly the Soil Wetness index?

"It represents, over a depth of about two meters, the state of the soil water reserve in relation to the useful reserve (for agriculture purposes)"

It represents the water state of the surface soil, not the water state of the water tables. The SWI is between 0 and 1, 0 being very dry soil where vegetation cannot subsist and 1 being the useful reserve, when the soil cannot absorb any more water.

Meteo France publishes publicly each year the monthly SWI on their website. The uniform SWI is the SWI calculated by the SIM model. More information can be found here in reference [8].

The data is a grid again. We adapt and use the same aggregating method as for the climatic data and built a monthly database between 2016 and 2020 by department for the SWI.

C. Modeling of the subsidence risk

As described in part A, insurers have little visibility of the subsidence risk. Indeed, due to the long CatNat system, insurance companies receive the subsidence claims on average 18 months after the event [5]. This makes it difficult to estimate the ultimate loss of a year of occurrence.

However, we do know at the end of a year of occurrence the climatic conditions and the severity of the drought on the French territory. The Météo France climatic data is available by the end of the first year of development. Hence, we can use external climatic data to compute a first estimate of the ultimate loss of the subsidence risk.

As presented before in part I.D, the number of claims seems to stabilize on the year of development four at maximum for each year of occurrence considered. Hence, our model will focus on predicting the number of claims for a year of occurrence at development four with external climatic data available at the end of the first year of development. We will then at the end add a tail coefficient to estimate the ultimate number of claims.

To avoid geospatial dependency issues, we model at the **scale of a department**. This means that one line is the data for one department of one year of occurrence in our dataset. This entails a low number of lines to train on: 95 if we train on one year of occurrence, 190 if we train on two years on occurrence. Furthermore, the variable to predict, the number of claims by department, is extremely unbalanced as it depends on where the drought was located during the year of occurrence. For these two reasons, we will have to be extremely careful during model training to avoid over-fitting. We will apply resampling and grid search methods to avoid such issues.

We have a lot of explicative variables, 88 in total, that will be described in part II.C.a. We will keep the most important ones through feature engineering.

Concerning modeling methods, we compare two different algorithms:

- **GLM:** a traditional model used in the Insurance sector, especially in risk scoring.
- **XGBOOST:** a popular model used for predictions especially in Kaggle competition. It is rumored to be efficient for such unbalanced datasets, so we will try to fit and use such model.

First, I will describe the available descriptive variables. Then, we will train and compare the two different algorithms and evaluate their robustness

a. Predictive variables

In total, there are **88 different explicative variables**. We will need to perform feature engineering to select the ones with the best predictive power. All 88 explicative variables are at the department level for one year of occurrence and have been extracted from the different databases described in the previous part. The variables and their original database can be found in the next table:

Label	Variable	Number of variables	Source
*tempAVG_*	Monthly average temperature	12	ECA&D
*tempMAX_*	Monthly maximum temperature	12	ECA&D
*tempMIN_*	Monthly sum of precipitation	12	ECA&D
*monthSPI_*	Monthly SPI	12	ECA&D
*monthSPEI_*	Monthly SPEI	12	ECA&D
*SPI_*	Seasonal SPI	4	ECA&D
*SPEI_*	Seasonal SPEI	4	ECA&D
*SWI_*	Monthly SWI	12	Météo France
nAlea	Average level of hazard of shrinkage-extension of clay soil	1	BRGM
expo	Exposure to risk: number of contracts	1	Allianz's portfolio data
Ratio_maison	Share of house in the portfolio	1	Allianz's portfolio data
hiver, printemps, ete, automne	Seasonal number of claims during the first year of development	4	Allianz's claim data
logchare	Log of the net charge of claims during the first year of development	1	Allianz's claim data

Table 5: Explicative variables and their original database

The explicative variables can be divided into 4 categories:

- **Climatic variables**: This category includes raw climatic variables such as minimum, maximum, average temperature and precipitations, but also indexes as the SPI, SPEI and SWI. For a year of occurrence, as seen in part I.B, we expect departments with abnormally low precipitation and high temperature to be more affected by drought. We compute the indexes at the monthly level, but also at the seasonal level. Indeed, as droughts occur each year at different month and our historical data being limited, we do not want the model to deduce that some months are more important in terms of drought than others. Using seasonal data could help keep the model flexible and avoid overfitting.
- **Fixed variables:** the variables that do not change from a year of occurrence to another. They are fixed characteristics of the department, such as the average risk of subsidence due to clay composition of the soil and the house percentage of the department in Allianz's portfolio. As described in part II.B.d, we expect a department with an important exposure to the shrinkage-extension of the soil to

be more affected by subsidence. We also expect a department with a lot of house in Allianz's portfolio to be more impacted by subsidence than a department with more buildings, as the foundations of houses are shallower.

- **Exposure:** This variable is the number of housing insurance contracts in each department. It is important to include it in the model as, by construction, there cannot be more claims than housing insurance contracts in a department, even if the department is severely affected by drought.
- **Claim variables**: at the end of the first year of development, we already have a few reported claims and their loss. The location of the claims could help indicate, or at least have an intuition, on which departments are the most affected by the annual drought. Hence, we create 5 variables out of the available data at the end of the first year of development. First, we use the log of the loss of the claims. Second, we record the numbers of claims declared in winter, summer, spring and fall to try to catch the seasonal effect of drought. We expect that the higher the numbers of the variables are, the more severe the year of occurrence will be.

Now that we have a first look at our descriptive variables and their relationship to the variable to predict, we will begin training our two models of interest, GLM and XGBOOST.

b. Generalized Linear Model

GLM, or the Generalized Linear Model, is a common tool in the insurance sector for pricing or risk scoring. The goal of GLM is to explain a random variable Y by p variables $X = (X_1, \dots, X_p)$. GLM is an extension of the linear model to accommodate for non-gaussian response distribution and transformations to linearity.

The theory behind GLM [20]

The assumptions of the model are the following:

1. Distributional assumptions: the conditional distribution of the response variable Y given X should be a member of the location-scale exponential family distribution and the different couple responses/covariate are independent with the same φ.
2. Structural assumption: the link function g (invertible) describes how the mean response $\mu = E(Y|X)$ is related to a linear combination of the predictors $\eta = X\beta$

$$\eta = g(\mu)$$

(x_i, y_i) follows a GLM if y_i given x_i are independent and

$$f(y_i|\theta_i, \phi) = \exp\left\{w_i \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}$$

$$\text{With } \mu_i = E(Y_i|X_i) = b'(\theta_i) \text{ and} g(\mu_i) = x_i{}'\beta$$

Hence

$$\mathrm{L}(\beta, \phi|\mathrm{y}, \mathrm{x}) = \prod_{i=1}^{n} \exp\left\{w_i \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}$$

Assuming ϕ is now, the estimation of the β parameters is done by maximum likelihood

$$\mathrm{LL}(\beta, \phi|\mathrm{y}, \mathrm{x})) = \sum_{i=1}^{n} \frac{w_i\left(y_i\theta_i - b(\theta_i)\right)}{\phi} + c(y_i, \phi)$$
$$\text{With } \mu_i = b'(\theta_i) \text{ and } g(\mu_i) = x_i{}'\beta.$$

Fitting GLM on our data

The characteristics of the GLM model that we will use are the following:

- **Variable to predict:** number of claims at year of development four for a department and for a year of occurrence
- **Descriptive variables**: I will describe shortly later the feature engineering process used to select the variables with the most predictive power out of the 88 previously described variables.
- **Functions:** We use a Poisson function to model a count variable using a log link function.
- **Offset:** We used the portfolio exposition as the offset, as to limit the number of claims to the number of contracts in the departments.

The process of feature engineering and model selection was quite long. For most models, especially the ones with the most variables, their predictive power was quite low. They tended to predict an abnormally large number of claims (more than 40 000!). The overall process can be found in annex 1.

Let us describe our best model optimizing the Root-Mean-Square Error (RMSE). We use the year of occurrence 2016 and 2017 as training sample (190 observations) and we use the following 10 variables:

- nAlear
- ratio_maison
- Seasonal SPEI of the year of occurrence
- Seasonal SPEI of the year before the year of occurrence

Using this model, all variables are significant at the 5% level. We get the following overall predictions for each year of occurrence found in the following table. It is important to keep in mind that only the years of occurrence 2016 and 2017 have reach development

four, meaning years of occurrence 2018, 2019 and 2020 are still incomplete in terms of number of claims reported.

Year of occurrence	Last year of development available	Number of claims recorded	Number of claims recorded at year of development 4	Number of claims predicted for year of development 4
2016	5	2 549	2 542	2 858
2017	4	3 768	3 768	3 447
2018	3	5 606	5 606	3 165
2019	2	3 044	3 044	4 132
2020	1	433	433	3 011

Table 6: Predicted versus observed number of claims by year of occurrence – Best GLM model

The results of our best GLM model are not so satisfactory. We know by experience (see part I.C) that 2018 is a particularly severe year for drought. Here, our model does not seem to be able to predict the **scale of severity** with the information that was given. The year of occurrence 2019, which is also quite severe but less than 2018, has a higher number of claims predicted than the year of occurrence 2018. 2018 even has a lower predicted number of claims than already reported number of claims at year of development three! The severity of year 2018 is also deemed similar to the severity of 2016 and 2017 with quite a close number of predicted claims.

It is possible for many reasons that GLM is not performing correctly here. First, our sample is quite small, and we have a lot of predictors. The sample is also highly unbalanced, based on the geographical characteristics of the annual drought on the French Metropolitan territory.

To deal with this situation, we can try to run and fit another algorithm that is known to be more efficient in these circumstances.

c. XGBoost

We train and test the XGBoost algorithm. XGboost is a recent algorithm released first in 2014 who has rapidly grown in popularity since.

Presentation and theory of XGBoost [21]

XGBoost stands for eXtreme Gradient Boosting and is an implementation of gradient boosted decision trees. It is conceived to have better performance and better speed than regular decision tree models.

But what is boosting and, furthermore, gradient boosting? **Boosting** is an ensemble technique. This means that, after the decision trees are run once, new models are added to correct the errors made by existing models. The models are added sequentially until there is no improvement observed. **Gradient Boosting** is a boosting method that uses a gradient descent algorithm to minimize the loss when adding new models. An example of the algorithm can be found in annex 2.

Fitting XGBoost: A first try

We first try to train the algorithm on the dataset 2016-2017 with quite standard parameters for XGBoost. We immediately face an issue: **overfitting**. As one can see in the following table and figure, it seems that XGBoost is overfitting on the train set: predictions for year of occurrence 2016-2017 are "perfect" meaning there are absolutely no prediction errors.

Year of occurrence	Last year of development available	Number of claims recorded	Number of claims recorded at year of development 4	Number of claims predicted for year of development 4
2016	5	2 549	2 542	2 542
2017	4	3 768	3 768	3768
2018	3	5 606	5 606	7 262
2019	2	3 044	3 044	7 114
2020	1	433	433	6 628

Table 7: First XGBoost predictions

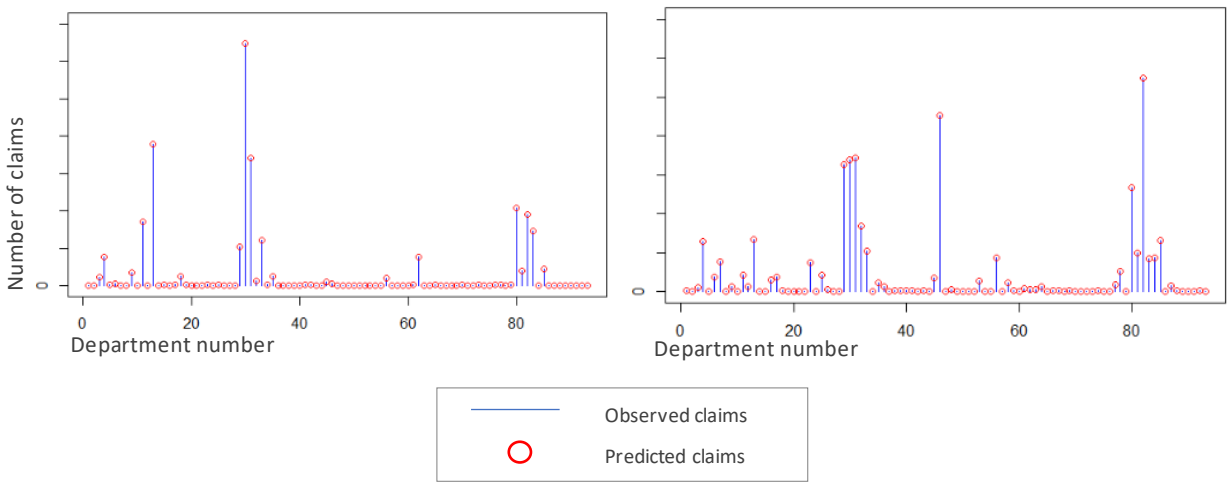

Figure 10: First XGBoost Predictions by department (2016, 2017)

For confidentiality reasons, we do not directly display the observed and predicted numbers of claims by department.

The reason we are facing overfitting might be our small training sample. XGBoost is an extremely powerful tool that can tend to overfit on such dataset. In our case, we absolutely want to avoid overfitting. Indeed, our years of occurrence are quite different from each other's in terms of months of occurrence of the drought and localization of the drought. Hence, we want to keep our model flexible to be able to predict for upcoming years of occurrence that could be quite different from the training set.

Fitting XGBoost: resampling and grid search

Avoiding overfitting is crucial in our case, as the years of occurrence have very different characteristics. It is quite hard to avoid, as we have very few observations and XGBoost is a very powerful tool. In this part, we will use resampling and grid-search to optimize XGBoost hyperparameters and avoid overfitting.

Let us first describe the sampling method. We use the years of occurrence 2016-2017 (190 observations in total). We draw randomly without replacement 70% for the train set and we keep the 30% as the test set. We renew this operation 200 times, which creates in total 200 different samples.

Alternatively, we create a hyper-parameter grid to conduct grid-search. XGBoost has many parameters to optimize. However, we only choose five parameters to tune to limit computing time. The parameter choice was made by experience but also through research on what parameters are most prone to lead to overfitting. Let us quickly describe what are the parameters chosen and how tuning them could help reduce overfitting:

- **Learning rate**: Slowing down the learning is quite a common method to avoid overfitting in such Machine Learning algorithms. The learning rate

will apply weights to the corrections of the new trees that are sequentially added to the model and control more efficiently the learning of the model to avoid overfitting.

- **Max depth**: The max depth of a tree. This is equivalent to the number of nodes of a tree. The higher the max depth, the more complicated the tree might be. To avoid overfitting, one tends to limit maximum depth to create simple and flexible trees.
- **Minimum child weight**: In XGBoost, for each node, a child weight is computed. Minimum child weight sets a minimum weight: if the child weight is underneath the threshold, the tree will stop at that node. This means that the higher the minimum child weight is set, the more complicated the tree might be, as the criteria to create a new node will be looser. To try to avoid overfitting, we can try to use smaller values of the parameter to simplify the model.
- **Column sample by tree**: The subsample ratio of columns when constructing each tree. This means that for each tree, different variables will be drafted if the parameter is fixed strictly below 1. If a predictor has a particularly important predictive power, this will help getting more diverse trees and avoiding relying too much on one variable for predictions.
- **Number of trees**: The total number of trees in the model. Optimizing this parameter is quite useful to have better performance as resampling and grid search will entail a significant computational time.

When combining the different hyperparameters together, we create in total 960 models. The different values of the parameters in the hyper grid can be found in the following table.

Number of trees	Learning rate	Max depths	Minimum child weight	Column sample by tree
10	0.01	1	1	0.8
40	0.05	3	3	0.9
60	0.1	5	5	1
100	0.3	7	7	

Table 8: Hyper-parameter grid

After running our 960 models through our 200 samples, we have to select the best models. We keep the ten models that on average perform the best on the 200 test samples. In other terms, we keep the ten models with the lowest average RMSE, our metric of interest, on the 200 test samples. The ten best models can be found below:

Ranking	Test RMSE	Train RMSE	Number of trees	Learning rate	Max depth	Minimum child weight	Column sample by tree
1	35.327	26.371	60	0.05	1	5	0.9
2	35.330	26.377	60	0.05	1	5	0.8
3	35.368	25.865	10	0.30	1	5	0.8
4	35.398	27.698	50	0.05	1	5	0.8
5	35.403	26.363	60	0.05	1	5	1.0
6	35.437	25.812	10	0.30	1	5	0.9
7	35.440	27.690	50	0.05	1	5	0.9
8	35.507	27.684	50	0.05	1	5	1.0
9	35.511	25.775	10	0.30	1	5	1.0
10	35.54291	25.365	60	0.05	1	3	0.8

Table 9: Top 10 models in term of RMSE

Looking at the table, one can see that the grid search recommends quite similar models. The recommended models are quite simple. All models have a depth of one, meaning their trees are only single nodes. The value of the learning rate is also low, usually below 0.1. The optimal number of trees varies between 10 and 60, however we will keep 60 to try to avoid bias. The column sample by tree of the best model is 90%, meaning that we do not keep all the predictive variables for each tree. As our training sample is quite small and the risk quite hard to model, it may be why such a simple model parameter set-up is recommended. The low number of trees and the low max depth will also help automatically select the best features for our model.

XGBoost: final results

Now, let us evaluate the predictions of our tuned XGBoost model. Because we do not use many trees in our algorithm, there is a part of randomness in the feature selection (column sample by tree is set to 90%). This means that the results vary from fit to fit of the XGBoost algorithm. To get more robust predictions, we train 100 fined-tuned XGBoost models and use the average of the predictions.

Let us first have a look at feature selection and feature importance. One of the advantages of XGBoost is the automatic feature selection. To measure the relative predictive power of a variable, we use the Gain metric. The Gain measures the relative contribution of the feature to the model by taking each feature's contribution for each tree in the model. The feature selection and feature gain can be found in the following figure:

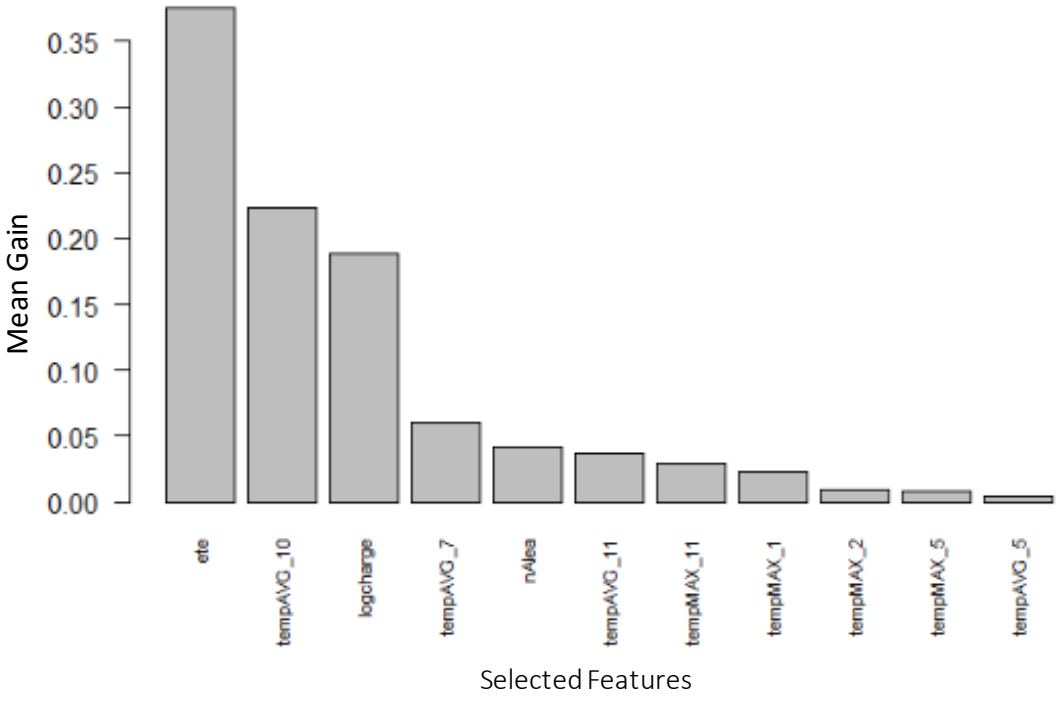

Figure 11: Mean gain of selected features of the tuned XGboost model

Looking at the graph, the training of XGBoost selected 11 out of the 88 variables. These variables, ordered decreasingly by mean Gain, are:

1. **Ete:** the number of claims recorded by the end of the first of development occurred in the summer of the year of occurrence.
2. **tempAVG_10:** The monthly average temperature of November for the department
3. **logcharge:** The log of the charge of the claims that have been reported at the end of the first year of development
4. **tempAVG_7**: the monthly average temperature of July for the department
5. **nAlea**: the average exposition to the shrinkage and extension of clay soil risk in the department
6. **tempAVG_11:** the monthly average temperature of November for the department
7. **tempMax_11:** the monthly maximum temperature of November for the department
8. **tempMax_1**: the monthly maximum temperature of January in the department
9. **tempMax_2**: the monthly maximum temperature of February in the department
10. **tempMax_5**: the monthly maximum temperature of April in the department
11. **tempAVG_5**: the monthly average temperature of April in the department

It is important to point out that:

- Predictive variables from the **claim data** are very important for the model, especially the log of the loss already recorded by the end of the first year of development and the number of claims occurred in the summer recorded by the end of the first year of development.
- Average and maximum **monthly temperatures**, without consideration for the season, are quite important to the model.
- No raw **precipitation variables nor precipitation indexes** were selected by the model. This is quite a surprising result and may induce complications later. Indeed, we know that the drought of 2018 for example is atypical due to the lack of precipitation in fall (see part I.C).

Let us now have a look at the overall predictions of the trained hyper-tuned XGBoost model by year of occurrence. For each year of occurrence, we sum the predictions for each department to get a predicted total number of claims. The results and comparisons can be found in the following table:

Year of occurrence	Last year of development available	Number of claims recorded	Number of claims recorded at year of development 4	Average Number of claims predicted for year of development 4
2016	5	2 549	2 542	2441.8
2017	4	3 768	3 768	3521.3
2018	3	5 606	5 606	6154.6
2019	2	3 044	3 044	4240.6
2020	1	433	433	3909.8

Table 10: Predicted versus observed claims by year of occurrence

A few interesting remarks are:

- Predictions for 2016 and 2017 are quite close to the observed numbers of year of development four without **overfitting** such as for the first XGBoost model. However, we can see that the predicted number of claims for the training sample is always below the actual reported number of claims. It is possible that our new model is a bit **optimistic**, meaning it predicts less claims than what is true.

- The **scale of severity** is conserved, contrarily to GLM: 2018 is by far the most serious year of occurrence in terms of predicted number of claims.

Let us now have a look at the detailed predictions by department. The following graphs show the predicted and the observed number of claims by department at year four of development for the occurrence years 2016 to 2020. One must keep in mind that only the years of occurrence 2016 and 2017 have had four years of development. Hence, it is difficult to evaluate the predictions of the years of occurrence of 2018, 2019 and 2020 with metrics as some of the claims have yet to be reported.

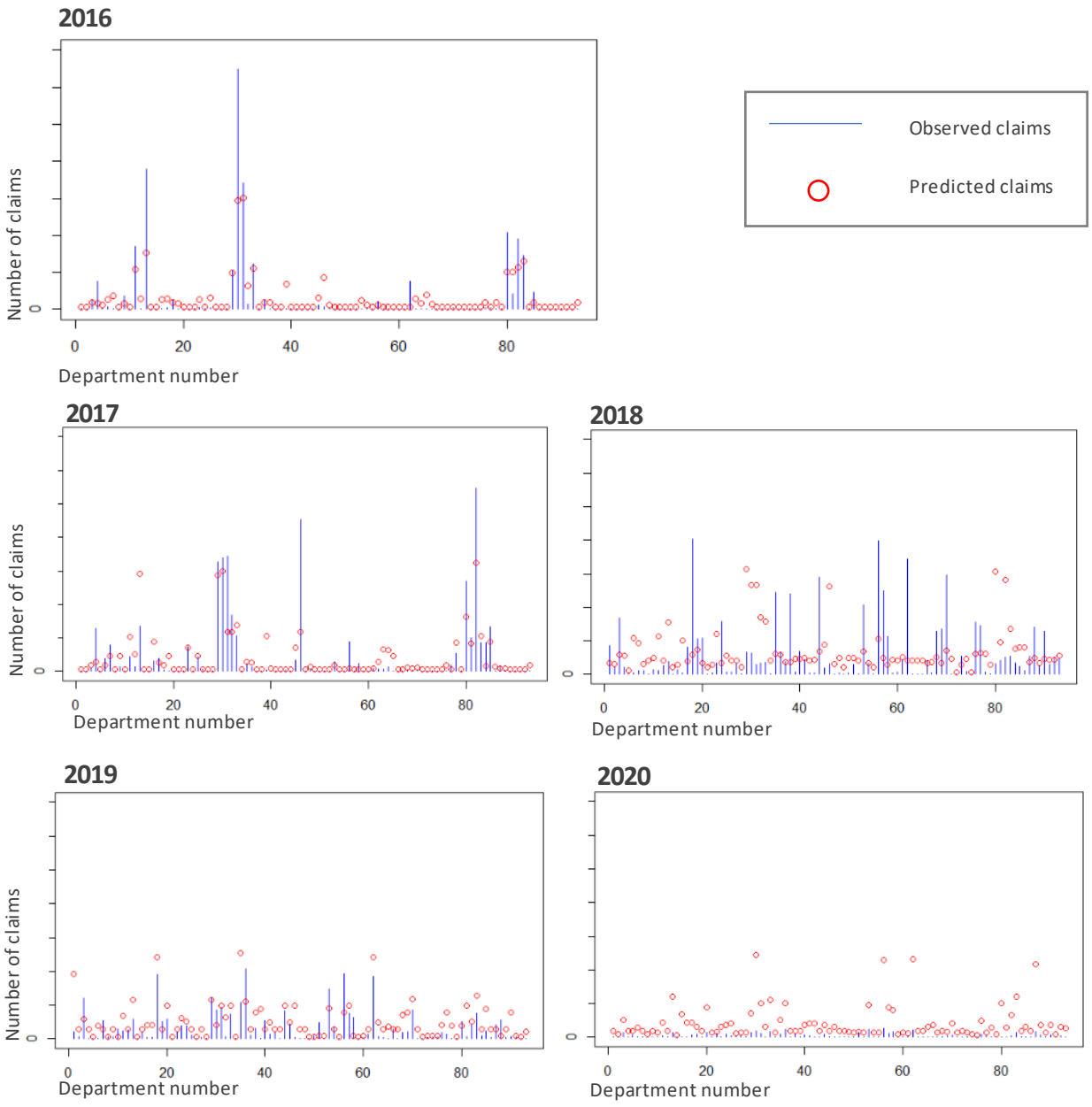

Figure 12: Predicted versus observed claims by department and year of occurrence

For confidentiality reasons, we do not directly display the observed and predicted numbers of claims by department.

First, it is important to point out that for the years 2016 and 2017, we seem to be able to avoid the **overfitting** effect: the predictions by department are not as perfect as before.

Second, the years of occurrence 2019 and 2020 have interesting predictions. Especially in 2019, the model can predict quite well most **sinistrality peaks** (departments 62 or 18 for examples) and when they are **no or only a few claims** (departments 21, 25...).

However, there are some issues with the **year of occurrence 2018**. The model cannot predict well department with a lot of claims or no claims. Looking at the graph, a lot of departments seem to have a predicted number of claims at around 25. Furthermore, sinistrality peaks are predicted for department in **the south east and south-west of France**: Haute-Garonne (31), Gard (30), Lot (46) and Tarn-et-Garonne (82). However, as seen in part X and X, these areas were not particularly affected by drought and subsidence during 2018. In fact, a surplus of rain was observed in the south during the fall 2018, protecting it from the severe drought, contrarily to the North-East/Center of France. It is possible that, **as our model does not consider any precipitation variables** or humidity index, it is not able to understand the event of the abnormal year of occurrence 2018 and is just inferring from **the southern temperature patterns**.

This issue definitively points out the limits of the model. As we have very few information and observations at the end of the first year of occurrence and we only train on two years of occurrence, we may lack historical train data to predict for atypical years. One way to counter this issue would be to retrain the model as soon as a year of occurrence reaches development four. By the end of 2021, we will be able to add an atypically year of occurrence (2018) to our training database and it will be interesting to see if it can improve the predictions.

One other issue of the model is that it might be slightly optimistic, meaning the predicted number of claims might be a bit below the actual number of reported claims (as seen for training years 2016 and 2017). We must keep this in mind when evaluating the total loss for each occurrence year.

Finally, one of the issues might also be the political aspect of the CatNat process. Claims are directly linked to a government decision as to whether to declare a municipality in a natural disaster state. It is possible that such decision does not entirely rely on meteorological or geotechnical criterion and that the model is lacking societal/political information to correctly predict.

d. Sensibility test of the XGBoost tuned model

To test the **robustness** of the results of the XGBoost tuned model, we try to train and predict with a new model. For this new model, we update the variables extracted from the claim data. Instead of computing the number of claims and the reported loss after one year of development, we update the values **to the second year of development**. As said before, the average time for a claim to be processed is two years, so by year of development two, it is possible that a lot more of information is available for the model. The question is: with a bit more claim information, can we predict **more accurately atypical years** such as 2018?

We use the parameters determined by the previous hyper-grid to avoid overfitting. We retrain the model on the updated database at year of development 2. Let us have a look at feature selection and feature importance:

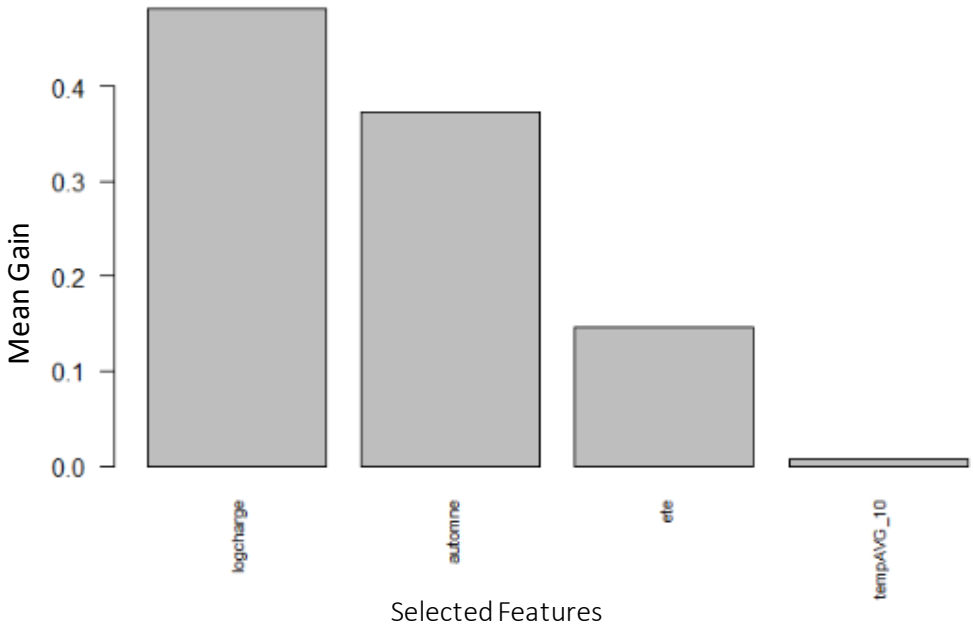

Figure 13: Feature selection and importance for the tuned XGBOOST model – Year of development two

Interestingly, the feature selection for the updated model is quite different from the previous optimized XGboost model. Now, XGboost selects only 4 variables and 3 of these 4 variables are extracted from the claim dataset:

1. **Logcharge**
2. **Automne**
3. **Ete**

4. **tempAVG_10**

As the first three variables are variables that have been **updated to the year of development two,** it is possible that the model has enough internal data as not to rely on the **climatic external data** as it did during the previous tuned XGBoost model.
Let us have a look at the overall predictions:

Year of occurrence	Last year of development available	Number of claims recorded	Number of claims recorded at year of development 4	Average Number of claims predicted for year of development 4
2016	5	2 549	2 542	2526.8
2017	4	3 768	3 768	3435.7
2018	3	5 606	5 606	7719.0
2019	2	3 044	3 044	4182.9

Table 11: Predicted Versus observed number of claims at year of development 4 by year of occurrence

It is important to point out that:

- The **scale of severity** is still conserved with 2018 being the most severe year of occurrence in terms of subsidence. It also seems that we were able to avoid overfitting.
- The predictions are more **pessimistic** for the year of occurrence **2018**. The model now predicts almost 8 000 claims at year of development four, around 2 000 more than for the previous model.
- The model might still **be optimistic**, as the predicted number of claims for year of occurrence 2016 and 2017 are still below the observed number of reported claims.
- We do not have results for the occurrence 2020, as it has still not reach development two.

We now have a look at the predictions by departments for the years of occurrence 2016 – 2019:

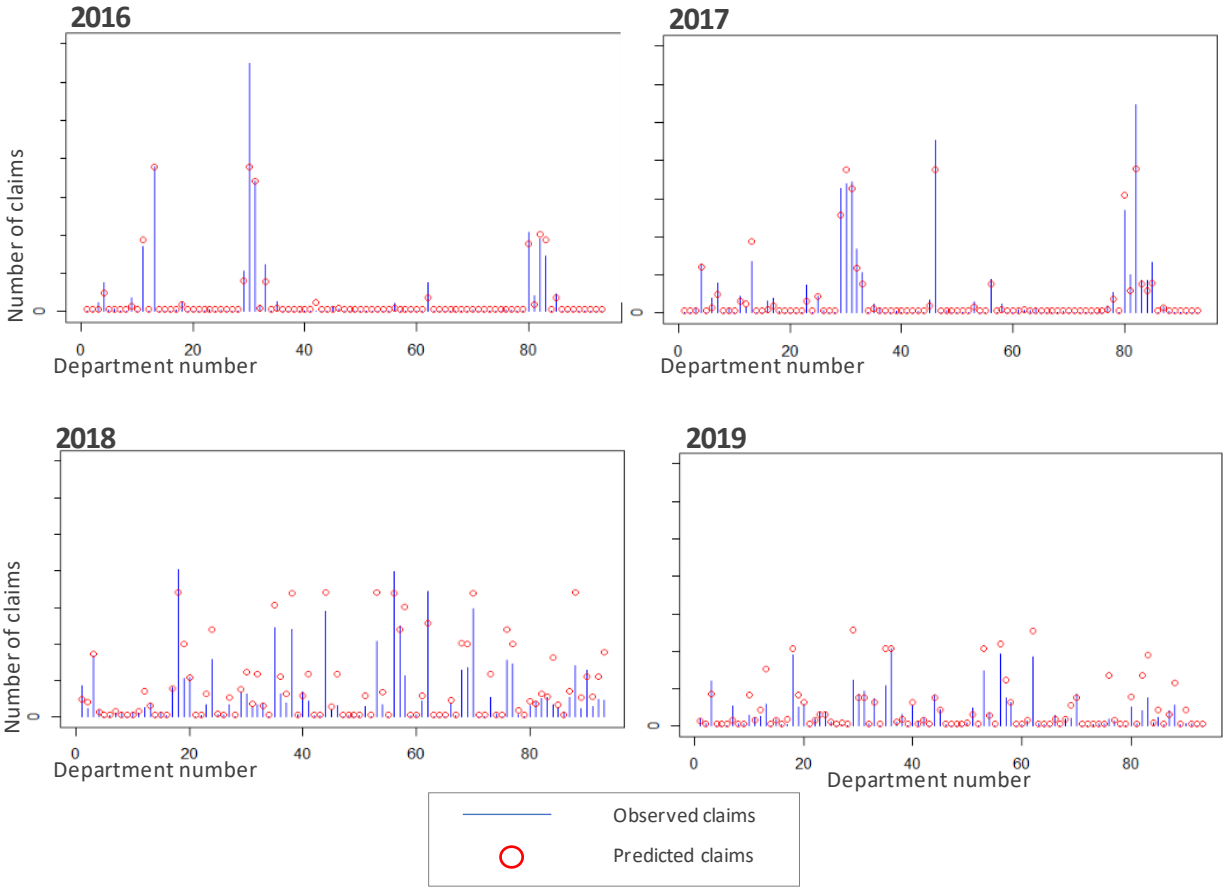

Figure 14: Observed vs predicted claims by department

The predictions by department for the year of occurrence 2018 seem to be more accurate than the previous XGBoost model. We do not observe peaks of claims for the southern departments anymore, meaning the errors in predictions might have come from a train bias in the temperature/precipitation variables. There also seems to be an improvement in predictions for departments with little to no claims, where it used to predict around 25 claims.

Concerning the occurrences 2016 and 2017 used to train and tune the model, there is not so much difference with the previous XGBoost model.

D. Model Prediction and business outlook

So far, we have modeled the subsidence and the drought risk. However, the real business-related question is not how many claims we observed, but how much the years of occurrence might cost to Allianz. This information will help the actuaries in the

department to reserve properly for the subsidence risk of our years of interest, but also for the years of occurrence to come if the model is to be updated. In this part, we will transform our model predictions to business output, as to be able to gauge the severity of each year of occurrence.

First, we will estimate the ultimate number of claims by year of occurrence. Then we will try to compute an estimate of the ultimate loss for each year of occurrence.

a. Ultimate number of claims and tail coefficients

During this project, we modeled the number of claims in a department for a year of occurrence at year of development four considering the data available at the end of year of development one. Rather than directly modeling the ultimate number of claims, we were constrained to make this decision, as we have no portfolio data before 2016.

We also made this decision as it seemed to make sense with the claim distribution. Indeed, over the years of development and for each year of occurrence, the number of claims seems to be a long-tailed distribution (see part I.D). During the first years of development, a lot of new claims are recorded, coming as soon as a natural disaster decree is published. Because of the decrees' publication, it is also during the first years that we can observe "shocks", an unusual large number of claims that made it difficult to use traditional reserving methods. Looking at the data, the threshold that was decided for "stability" is year of development 4. This means that we consider that after development four there are no more shock observed and the number of claims is increasing slowly. We consider as the ultimate number of claims the number of claims for year of development 20, as it the oldest our historical data can go. Looking at year of occurrence 2001 in the claim triangle (annex 3), the number of claims in the last years of development is stable, with no increase whatsoever. Hence, this seems like a reasonable assumption.

Knowing this, we can model the ultimate number of claims using our historical claim data starting from 2001. We assume that the number of claims over the years of development is a left long-tailed distribution and that no shock can happen after year 4 of development.

Let us have a first look at the "tail" and the overall number of claims left to be reported at the end of the year of development 4 in the next table. One must keep in mind that the most recent years (above 2010) have had few years of development since year of development 4, hence it is normal to observe low number of claims left to be reported.

At year of development 4	Number of claims left to be reported	Total number of claims reported	Tail pourcentage
Moyenne 2001-2016	40.3	1511.9	2,7%
2001	10	230	4,3%
2002	63	1791	3,7%
2003	307	9832	3,3%
2004	55	523	11,0%
2005	76	1786	4,4%
2006	12	464	2,6%
2007	19	553	3,5%
2008	6	426	1,4%
2009	6	884	0,7%
2010	6	205	3,0%
2011	22	4033	0,6%
2012	16	1213	1,4%
2013	2	126	1,4%
2014	0	88	0,0%
2015	4	521	0,7%

Table 12: Number of claims that were yet to be reported at the end of year 4 of development, Subsidence historical claim data (2001-2015)

On average, 40.3 claims are yet to be reported at the end of year 4 of development, however this number varies greatly based on the severity of the drought that occurred during the year. To understand better how to compute the ultimate number of claims, we have a look at the tail coefficient. The tail coefficient is the number of claims yet to be reported at year of development 4 divided by the total number of claims reported in 2020. Looking at the different years of occurrence, the tail coefficient seems to be quite stable, average at 2.7% and varying from 0% to 4.3%. The only disparity is the year of occurrence 2004 with a tail coefficient of 11%. As this is one of the oldest years of occurrence in our database and we expect the CatNat process to have improve since, we are not taking this year of occurrence into account for the analysis.

As our analysis reinforce our previous assumption, we can now establish a method to compute the ultimate number of claims:

1. **Compute the tail coefficient** for each year of occurrence using the results of the chain ladder method on the number of claims (see annex 3). With this method, we use Chain Ladder's factors of development to understand the evolution of the tail through the years of development. We divide the difference between the chain

ladder ultimate and the number of claims at development four by the ultimate to get the tail coefficient.

2. Appling the tail coefficient on the predicted number of claims for both XGBoost models and **computing the final ultimate number of claims** (different from the chain ladder ultimate).

The results of the method can be found in the following table:

Year of occurrence	Predited number of claims - XGBOOST DEV 1	Predited number of claims - XGBOOST DEV 2	Tail coefficient	Ultimate number of claims – DEV 1	Ultimate number of claims – DEV 2
2016	2441.8	2526.8	1.4%	2476.0	2562.2
2017	3521.3	3435.7	2.8%	3619.9	3531.9
2018	6154.6	7719.0	2.8%	6326.9	7935.1
2019	4240.6	4182.9	2.8%	4358.7	4300.0
2020	3909.8		2.8%	4019.3	

Table 13: Ultimate number of claims by year of occurrence and by chosen XGBOOST model

The difference between the predicted number of claims at year 4 and the ultimate predicted of claims is not so significant. This is compliant to our expectations and our previous analysis as we are dealing with a long-tailed distribution.

b. Estimation of the ultimate loss for each year of occurrence

For the reserving department, it is important to be able to assess the cost of a year of occurrence to reserve the proper amount. So far, we have only estimated the ultimate number of claims rather than their cost. Now, we will focus on the cost of a year of occurrence to be able to produce an estimate and present it to the actuaries of the Reserving department. A focus year will be 2018 as it might be a very costly year for the insurance company and its real cost is still unclear today.

To estimate the total loss for a year of occurrence, we will use the average loss for one claim. For each year of occurrence, we divide the total loss by the total number of claims. As the average loss of a claim varies greatly form one year of development to another, we use the ultimate average loss computed through chain-ladder (see Annex). As it seems to be quite an uncertain estimate, we rather create a value range of the average cost of a claim for each year of occurrence. Using the Chain ladder average loss

estimate, we create a lower range value, removing 20% of the estimate, and a higher range value, adding 20% of the estimate. The final average loss by claim estimates by year of occurrence can be found below:

Year of occurrence	Average claim loss - Lower	Average claim loss - Intermediate	Average claim loss - Higher
2016	37 749.4	47 186.8	56 624.2
2017	29 367.9	36 709.9	44 051.9
2018	24 032.4	30 040.5	36 048.6
2019	21 641.3	27 051.6	32 421.9
2020	18 230.9	22 788.6	27 346.3

Table 14: Average loss for one claim depending on the year of occurrence (in euro)

Using the average loss by claim, we can compute an estimate of the ultimate total loss for a year of occurrence by multiplying it by the estimated ultimate number of claims. We run this computation twice: for the optimized XGBoost model with the data from the first year of development and the XGBoost model with the data from the second year of development. The results can be found in the two tables below:

Year of occurrence	Ultimate number of claims	Ultimate Loss - Lower	Ultimate Loss - Intermediate	Ultimate Loss - Higher
2016	2476	93 467 514	116 834 517	140 201 519
2017	3620	106 311 798	132 889 838	159 467 878
2018	6327	152 052 995	190 066 244	228 079 492
2019	4359	94 334 427	117 917 924	141 327 062
2020	4019	73 269 987	91 587 383	109 904 780

Table 15: estimate for total loss ultimate by year of occurrence (in euro) – XGBOOST first year of development

Year of occurrence	Predited number of claims	Ultimate Loss - Lower	Ultimate Loss - intermediate	Ultimate Loss - Higher
2016	2527	95 392 734	110 241 044	143 089 353
2017	3436	100 908 104	126 135 216	151 362 328
2018	7719	180 506 096	231 882 620	278 259 143
2019	4183	90 525 558	113 156 843	135 620 808

Table 16: estimate for total loss ultimate by year of occurrence (in euro) – XGBOOST second year of development

The results are a construction from our previous predictions, so they are quite **similar** to the results that we had in terms of number of claims. For example, the total loss for the year of occurrence 2018 might be underestimated by the first XGBoost model, as this model does not consider precipitations. The second model gives an estimate for the ultimate total loss of around 63 million. The higher range of estimate is around 280 million. Compared to similar projects that have been done in other departments of the company, those results are slightly lower, meaning more optimistic in term of loss. For example, the technical direction estimated the ultimate loss for 2018 as around 285 million.

As we do not have so much historical data to compare to, it will be interesting to see how the situation evolves in the upcoming years. It will also be interesting to update the model as soon as the years of occurrence reach the fourth year of development. Indeed, one of the main issues of our model is that our train sample is very limited. Furthermore, two years of drought is actually very little information as drought are very different in terms of locations and time depending on the year of occurrence (see part I.C). Hence, expanding the training set will enable for a more flexible model, that may be able to comprehend abnormal years in terms of subsidence (such as 2018).

CONCLUSION

Since 2015, severe droughts have entailed a significant increase not only in terms of number of claims but also in the severity of the subsidence claims. This change have reinforced the need of insurers for modeling and predicting drought and subsidence. However, droughts are caused by complex climatic and geotechnical mechanisms that can be hard to model. Moreover, the French Natural Disaster Governmental process can lead to important delays for the insurers to process the subsidence claims. All these reasons lead to difficulties using regular reserving methods such as the Chain-Ladder method to reserve for each year of occurrence. This Master thesis proposed two different models to estimate the ultimate number of claims of a year of occurrence by the 31st December of the occurrence using external climatic data. By construction, we can deduce the estimated ultimate financial loss for each year of occurrence between 2016 and 2020.

Overall, the results of the tuned XGBoost model can give a first estimate of the severity of a year of occurrence with only data from the first year of development. However, the results are best with a model updated with data from the second year of occurrence. For example, for the year of occurrence 2020, we predicted an ultimate financial loss between 17 million and 26 million for 1961 claims. It is less than for the year of occurrence 2018 for which we predict between 40 and 75 million for between around 3200 and 4000 claims. Knowing that 2018 is a particularly severe year for drought, it seems that the model does gauge severity. However, comparing with our training sample and other projects of the company, it is possible that the model is optimistic, meaning that it predicts less claims than what is true. As most of the years of occurrence have not yet reach the year of development four, it is hard to evaluate more thoroughly the model. Overall, even with climatic and geotechnical external data, it is quite hard to model the subsidence risk. This may be partly caused by the CatNat governmental process that also depends on political and socio-economical information that the model does not have access to.

To improve the model, a first step would be to expand the historical database to be able to train the models on a bigger set of observations. For example, when the occurrence 2018 reaches development 4 in 2022, its integration to the model could be beneficial as it will include different precipitation patterns than in 2017 and 2016 within the model. Having reach development 4 on most of the years of occurrence will also help evaluating and adjusting the future predictions. Another way to improve the process would be to integrate variables about the French natural disaster governmental process, such as the departments that have had historically a significant number of municipalities declared in a natural disaster state. However, the criteria of the CatNat process have changed rapidly over the past decades and it might keep changing in the future with the increase in the occurrence and the severity of droughts.

REFERENCES

1. GIEC's Sixth Assessment Report, 2021, www.ipcc.ch/assessment-report/ar6/
2. A. CHARPENTIER, L. BARRY & M. JAMES, 2021, Geneva Papers on Risks & Insurance, Insurance against natural catastrophes: balancing actuarial fairness and social solidarity
3. CCR, 2011, Le Régime d'indemnisation des catastrophes naturelles
4. Ministère de l'Ecologie, du Développement Durable et de l'Energie, L'Aléa retrait-gonflement des argiles. http://www.georisques.gouv.fr/dossiers/alea-retrait-gonflement-des-argiles#/.
5. Mission des Risques Naturels (MRN), 2019, Lettre d'information de la mission des risques naturels
6. JF. SCHULTE, 2016, Modélisation du risque subsidence en France métropolitaine, Master thesis
7. A. CHARPENTIER, M. JAMES & H. HANI, 2021, arXiv preprint, Predicting Drought and Subsidence Risks in France
8. Météo France dans le dispositif CatNat des sécheresses, http://www.meteofrance.fr/documents/10192/79826318/M%C3%A9t%C3%A9o-France+dans+le+dispositif+CATNAT+s%C3%A9cheresse
9. A. IGLESIAS, D. ASSIMACOPOULOS & H. VAN LANEN, 2019, editors, Drought: Science And Policy, Wiley-BlackBell
10. Mission des Risques Naturels (MRN), 2021, Bilan des principaux évènements cat-clim. Lettre d'information, (35), 2021
11. Rapport sécheresse 2018, CCR, https://catastrophes-naturelles.ccr.fr/-/secheresse-2018-en-france
12. Rapport sécheresse 2016, CCR, https://catastrophes-naturelles.ccr.fr/-/secheresse-2016-en-france
13. Rapport sécheresse 2017, CCR, https://catastrophes-naturelles.ccr.fr/-/002088_secheresse-2017-en-france?inheritRedirect=true&redirect=%2Frecherche%3Fq%3Ds%25C3%25A9cheresse%2B2017
14. Rapport sécheresse 2019, CCR, https://catastrophes-naturelles.ccr.fr/-/002111_secheresse-2019
15. Rapport sécheresse 2020, CCR, https://catastrophes-naturelles.ccr.fr/-/002121_sech_2020
16. M. LANGUILLE, 2020, Elaboration d'un modèle de projection de charge ligne à ligne pour le provisionnement des sinistres corporels, Master Thesis
17. E. Arnaud, 2016, Modélisation du risque sécheresse en France, Master Thesis
18. European Climate Assessment & Dataset. E-OBS gridded dataset. http://www.ecad.eu
19. BGRM, Soil composition data access, https://www.georisques.gouv.fr/donnees/bases-de-donnees/retrait-gonflement-des-argiles
20. C. THOMAS, 2021, Scoring, class materials
21. XGBoost documentation, https://xgboost.readthedocs.io/en/latest/
22. R Studio, https://www.rstudio.com/

APPENDIX

Annex 1: Model selection for GLM – list of models

Model name	Training set	Variables included	Number of variables
"first try" model	Occurrence 2016	All 88 described variables	88
"Monthly" model	Occurrence 2016	Monthly SPI and SPEI + fixed variables	54
"Seasonal" model	Occurrence 2016	Seasonal SPI and SPEI + fixed variables	20
"first try" model	Occurrence 2016 and occurrence 2017	All 88 described variables	88
"Monthly" model	Occurrence 2016 and occurrence 2017	Monthly SPI and SPEI + fixed variables	54
"Seasonal" model	Occurrence 2016 and occurrence 2017	Seasonal SPI and SPEI + fixed variables	20
"Multiplicative" model	Occurrence 2016 and occurrence 2017	Seasonal SPI and SPEI + fixed variables + multiplication between consecutive seasonal SPI and SPEI	20
Simple model 1	Occurrence 2016 and occurrence 2017	Seasonal SPI + fixed variables	10
Climatic model	Occurrence 2016 and occurrence 2017	Raw monthly climatic data of the year of occurence + fixed variables	39
Best model (simple model 2)	Occurrence 2016 and occurrence 2017	Seasonal SPEI + fixed variables	10

Annex 2: Standard unregularized XGBoost algorithm [21]

Input: training set $\{(x_i, y_i)\}_{i=1}^{N}$, a differentiable loss function $L(y, F(x))$, a number of weak learners M and a learning rate α.

Algorithm:

1. Initialize model with a constant value:

$$\hat{f}_{(0)}(x) = argmin_{\theta} \sum_{i=1}^{N} L(y_i, \theta)$$

2. For $m = 1$ to M:
   1. Compute the "gradients" and "hessians":

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)}$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)}$$

   2. Fit a base learner (or weak learner, e.g. tree) using the training set $\left\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^{N}$ by solving the optimization problem below:

$$\hat{\phi}_m = argmin_{\phi \epsilon \Phi} \sum_{i=1}^{N} \frac{1}{2} \hat{h}_m(x_i) \left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)\right]^2$$

$$\hat{f}_m(x) = \alpha \hat{\phi}_m(x)$$

   3. Update the model:

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x)$$

3. **Output** $\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x)$

Annex 3: Chain ladder on Number of claims

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2 001	140	205	213	219	226	226	226	227	227	228	228	229	229	229	229	229	229	229	229	229
2 002	114	802	1 567	1 726	1 753	1 756	1 784	1 786	1 789	1 789	1 791	1 792	1 792	1 792	1 792	1 792	1 792	1 792	1 792	1 792
2 003	1 052	3 910	7 528	9 514	9 730	9 804	9 815	9 820	9 828	9 829	9 832	9 832	9 832	9 832	9 833	9 833	9 833	9 834	9 834	9 834
2 004	404	451	461	464	506	515	521	521	521	521	521	521	521	521	521	521	521	521	521	521
2 005	226	294	304	1 707	1 750	1 767	1 776	1 778	1 780	1 781	1 782	1 784	1 784	1 784	1 784	1 786	1 786	1 786	1 786	1 786
2 006	85	95	427	452	455	458	463	464	464	464	464	464	464	464	464	464	464	464	464	464
2 007	31	430	515	535	549	550	552	552	553	553	553	554	554	554	554	554	554	554	554	554
2 008	218	354	409	421	425	425	425	425	426	426	426	426	426	426	426	426	426	426	426	426
2 009	89	110	852	876	878	881	881	881	881	882	882	882	882	882	882	882	882	882	882	882
2 010	37	152	195	198	203	203	203	203	203	203	205	205	205	205	205	205	205	205	205	205
2 011	99	3 414	3 960	4 010	4 021	4 026	4 029	4 030	4 032	4 032	4 034	4 035	4 035	4 035	4 035	4 035	4 035	4 037	4 037	4 037
2 012	345	1 071	1 185	1 196	1 205	1 208	1 211	1 213	1 213	1 213	1 213	1 214	1 214	1 214	1 214	1 214	1 214	1 214	1 214	1 214
2 013	78	116	122	124	124	126	126	126	126	126	126	126	126	126	126	126	126	126	126	126
2 014	21	60	82	88	88	88	88	88	88	88	88	88	88	88	88	88	88	88	88	88
2 015	58	328	493	517	520	520	522	522	522	522	524	524	524	524	524	524	524	524	524	524
2 016	115	2 177	2 527	2 557	2 564	2 579	2 586	2 588	2 590	2 592	2 592	2 594	2 594	2 594	2 594	2 594	2 594	2 594	2 594	2 594
2 017	260	3 241	3 737	3 792	3 859	3 880	3 893	3 895	3 899	3 899	3 901	3 903	3 903	3 903	3 903	3 903	3 903	3 903	3 903	3 903
2 018	480	5 712	6 082	7 168	7 293	7 334	7 357	7 363	7 369	7 371	7 373	7 375	7 375	7 375	7 377	7 377	7 377	7 379	7 379	7 379
2 019	664	3 460	4 838	5 702	5 800	5 832	5 852	5 856	5 860	5 862	5 864	5 866	5 866	5 866	5 866	5 868	5 868	5 868	5 868	5 868
2 020	535	2 825	3 948	4 654	4 733	4 760	4 777	4 779	4 783	4 785	4 787	4 787	4 787	4 787	4 789	4 789	4 789	4 789	4 789	4 789

Annex 4: Average loss by claim triangle (Chainladder)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2 001	2 847	2 492	3 347	4 278	5 249	6 136	6 097	6 626	6 590	6 907	7 203	7 118	7 546	7 266	7 266	7 266	7 266	7 266	7 266	7 266
2 002	2 228	6 826	9 013	10 410	10 681	11 483	11 867	11 775	11 541	11 820	11 944	12 218	12 029	11 966	11 926	12 043	12 003	12 336	12 336	12 336
2 003	4 976	6 235	6 985	7 579	8 500	9 077	9 395	9 701	9 828	9 781	9 877	9 947	10 010	9 972	9 913	9 869	9 854	9 833	9 834	9 834
2 004	2 245	668	370	311	2 750	3 058	3 471	3 430	3 765	3 766	3 711	3 711	3 693	3 693	3 693	3 693	3 693	3 701	3 701	3 701
2 005	2 153	982	828	10 308	13 425	14 351	14 925	15 246	15 178	14 816	14 971	14 940	14 942	15 128	15 010	15 179	15 150	15 206	15 207	15 207
2 006	4 228	2 546	9 592	14 907	16 473	16 304	15 544	15 618	15 846	15 665	16 119	16 047	16 396	15 865	16 481	16 489	16 458	16 519	16 520	16 520
2 007	1 735	8 749	16 302	22 275	22 362	21 953	23 652	23 303	23 152	22 574	23 575	24 716	24 240	24 168	24 069	24 081	24 037	24 125	24 126	24 126
2 008	3 708	8 078	13 418	14 765	13 482	13 192	13 217	13 136	12 647	11 852	11 822	11 822	12 305	12 266	12 216	12 222	12 199	12 244	12 244	12 244
2 009	6 795	8 544	26 579	28 467	26 591	26 588	26 700	25 971	26 007	26 039	27 410	27 878	27 951	27 863	27 749	27 762	27 710	27 812	27 813	27 813
2 010	2 517	23 266	24 265	22 053	22 932	23 383	23 177	22 766	22 870	22 870	22 697	22 931	22 991	22 918	22 824	22 835	22 793	22 876	22 878	22 878
2 011	3 736	11 611	19 269	23 317	24 821	25 498	25 296	25 258	25 462	25 581	25 982	26 248	26 318	26 235	26 128	26 141	26 091	26 187	26 189	26 189
2 012	7 932	9 499	10 591	11 334	13 735	13 349	13 576	14 183	14 593	14 553	14 782	14 934	14 974	14 926	14 865	14 871	14 843	14 898	14 900	14 900
2 013	7 912	6 863	24 470	22 613	24 570	25 141	28 308	27 583	27 747	27 674	28 108	28 397	28 473	28 382	28 266	28 280	28 227	28 329	28 332	28 332
2 014	2 732	18 242	47 978	37 324	37 907	32 148	32 111	32 513	32 706	32 620	33 131	33 472	33 562	33 455	33 317	33 332	33 271	33 394	33 395	33 395
2 015	10 001	8 127	15 379	21 330	23 506	24 604	25 135	25 450	25 601	25 535	25 934	26 201	26 271	26 186	26 080	26 093	26 044	26 140	26 141	26 141
2 016	17 532	12 245	25 830	37 820	42 644	44 413	45 371	45 941	46 213	46 091	46 814	47 295	47 421	47 269	47 077	47 099	47 012	47 184	47 186	47 186
2 017	5 932	6 851	23 691	30 315	33 176	34 552	35 298	35 741	35 952	35 859	36 419	36 794	36 892	36 775	36 624	36 642	36 573	36 708	36 710	36 710
2 018	5 661	10 577	21 393	24 808	27 150	28 275	28 885	29 248	29 422	29 344	29 804	30 110	30 190	30 094	29 970	29 985	29 929	30 038	30 040	30 040
2 019	6 046	11 234	19 262	22 336	24 444	25 458	26 008	26 334	26 490	26 420	26 834	27 110	27 182	27 096	26 984	26 998	26 948	27 046	27 048	27 048
2 020	5 851	9 465	16 228	18 819	20 594	21 449	21 912	22 187	22 318	22 259	22 609	22 841	22 903	22 829	22 735	22 747	22 704	22 788	22 788	22 788