Subsidence and household insurances in France : geolocated data and insurability

Pierre Chatelain, Stéphane Loisel

Subsidence and household insurances in France : geolocated data and insurability

P. Chatelain, S. Loisel

Univ Lyon, UCBL, ISFA LSAF EA2429, F-69007, Lyon, France

Abstract

The insurability of natural disasters has always been an issue faced by the insurers, states, and insured persons. In France, the insurer and the legislator are concerned about the subsidence risks due to several consecutive dry years. More and more open data are provided in France, which allows insurers by geolocating their portfolio to have better knowledge. This knowledge plus the increase in subsidence risks query the insurability of the subsidence risk. Using mostly GLMs, the most common models used in France, this paper shows the improvement of the knowledge subsidence risks. The results bring to the fore the importance of legislative control and the recently enforced new CatNat program, leading authors to question the CatNat fee stagnation.

Keywords: Subsidence, actuarial pricing, reserving.

1. Introduction

France has a particular reinsurance program for natural disasters, especially non-life insurance. In France, the law about climatic catastrophes - *CatNat* - 13 July 1982 - L. 125-1 of *Code Des Assurances*, defines the effect of natural disaster as *"the direct and uninsurable property damage being the main cause of the abnormal intensity of a natural agent when the usual measures to be taken to prevent such damage could not prevent their occurrence or could not be taken."* Natural disasters abiding (drought, floods, earthquake, hurricanes ...) to this definition are considered Natural Catastrophe (CatNat). Once a municipality has declared CatNat, the claim's indemnities are compulsory for goods that are insured against fire damage. The legislator modified slightly the legislation the 28 December 2021 proving its long-term robustness. Developed first for floods, damages due to drought (such as clay shrinkage/subsidence) started to be

Email address: pierre.chatelain.act@gmail.com (P. Chatelain, S. Loisel)

taken into account in 1989. If the flood damage corresponds to 46 % of the CatNat damage, subsidence risks have increased in recent years (especially from 2016 to 2019). Insurers and the legislator are starting to consider seriously the increase in the frequency of subsidence (more globally natural disasters) due to first urbanization and secondly to climate change.

France and subsidence. Subsidence damages on buildings come from a low sinking of an area, mainly due to meteorological factors. In recent years, the frequency has increased. Uncountable papers explained the relationship between climate change and droughts. For instance, the report [11] highlights the fact that the meteorological drought evolution differs depending on the countries studied. More specifically, drought frequency, drought severity, and duration increase in southern France ([17]). Droughts are studied using different meteorological indexes; [10] study the standardized precipitation index (SPI), the standardized precipitation evapotranspiration index (SPEI), and the self-calibrated Palmer drought severity index scPDSI. In this paper, the indicators tested are the SPI, the Standardized Soil Wetness Index (SSWI), and the USA Reclamation Droughts Index (RDI) to model this risk. Regarding the actuarial, hydrological, and subsidence literature, Charpentier et al., 2021 [5] provides a complete overview. In short, the subsidence risk is a difficult risk to modelize with the current underwriting knowledge; more and better data are needed. Recently, more and more insurers are reducing the accessibility of new contracts due to subsidence presence. The recent drought of 2022 was asses as the second-worst drought in France after the 2003 one.

Open data et geolocation. According to the data.europa.eu, French open data are one of the leading European countries in 2021 in terms of accessibility and transparency. All this information allows the different actors to geolocate the building using the address, and have access to a subsidence vulnerability map or information on the building (number of floors, the construction period, the surface, the vegetated surface). In this paper, a historical household insurance portfolio is used and geolocated thanks to a data provider. The latter also provides around 60 variables on each building, as well as meteorological variables.

Contributions. This paper highlights several contributions. This article shows how to model subsidence risks for household insurance in France with the maximum data at disposable thanks to geolocated buildings. The models on the CatNat declaration at the municipality is highly improved adding meteorological information and aggregated information from the building level. Thanks to the latter, the results of the modeling of the frequency conditionally to a CatNat declaration show that the new model is more performant and segmenting than a model using only underwriting and reserving variables. Even if the claims' development censors part of the information, our model outperforms the traditional cost models through variables on the building and urbanization. Looking through the uses of these models *e.g.*

reserving, prevention, this paper shows that the performance gain of these models may influence the insurability of subsidence risks.

Section 2 lists all the data used with their particularity to fully understand all the models on the expected cost of subsidence proposed in Section 3. Finally, Section 4 discusses the problem of insurability of subsidence risks due to external data integration and some idea to adapt the French CatNat program.

This paper is based on a unique insurer's portfolio, which has allowed us to present these results. Moreover, one other insurer portfolio has used the same methodology and led to the same results (variables' selection and performance). All the numbers given are modified to anonymize the results or will be mentioned otherwise.

2. Data and subsidence

The available information on a contract usually stems from the underwriting process (see Subsection 2.1) and from some external data sources (see Subsection 2.3). However, knowing the exact geolocation, the buildings' information can be added, as detailed in Subsection 2.2. Data are gathered by a data provider which has created a database for insurance purposes in France. For other damage coverages, this new information has proven its value on the risks' knowledge. This data provider also hands meteorological indexes. Especially, droughts are detected through several meteorological indexes such as SPI, SSWI, or RDI which are detailed in Section 2.4. Finally, each insurer has his own portfolio's particularity, being detailed in Subsection 2.5.

2.1. Underwriting data

The portfolio considered is taken from a French insurer for MRH insurance (Multi-Peril Housing). The coverage of subsidence is compulsory only for owner-occupant or owner-non-occupant insurance contracts. This work will focus on both insurance contracts from 2015 to 2020 on the French mainland territory, excluding Corsica. Variables at the disposable are stemming from the underwriting process. The most relevant variables available are the occupant's age, the surface insured, the number of rooms, the period of construction, the personal property insured, the reconstruction value, and the type of contract (owner-occupant or owner-non-occupant) taken out. Each information is taken from the last update of the contracts' database in April 2021. The quality of the information is excellent for most of the variables. The information on the claim is the payment *pay*, the reserve *res* and financial recourse, and if the indemnity process is closed or open. We denote the *cost* at the date t as the sum of *pay* and *res* at the date t. We let aside the recourse, which is negligible in number

and in amount.The reserve process for a claim S is as follows :

$$res(S,t) = \begin{cases} 0 & \text{if no claims declared or closed,} \\ \hat{S} - pay(S,t) & \text{if else a claim } S \text{ evaluated and approved by an insurance expert,} \\ 20000 & \text{if else claim declared \& municipality declared a CatNat,} \\ 20 & \text{if else claim declared \& municipality has yet to declare a CatNat,} \\ 0 & \text{otherwise.} \end{cases}$$

To properly model claims costs, claims triangles from 2001 to 2020 about the number of claims, the payment, and the cost of subsidence were given. Different methods of reserving were used. The number of claims triangle development is better developed using GLM as proposed by [16] where the negative binomial is the better-suited distribution (as for [5]). Then, we develop the mean cost of subsidence damage using the Mack stochastic model [12]. Several reasons explain the choice to use the mean cost and not the complete reserve or payment triangle; the evaluation of the *res* of subsidence is quite erratic, with negative increments at some periods. Plus, in recent years starting from 2017, the payment does not provide sufficient information. The reserve is informative only once the insurance expert adjusted it. The open claims' costs are developed using the following factor Dev^{open}_{factor} for the J-the year of development :

$$Dev^{open}_{factor}(J) = \frac{CM(J)}{CM^{open}(J)} \frac{f_{Mack}(J)}{Prop_{open}(J)} \tag{1}$$

where $CM(J)$ is the mean cost of all claims, $CM^{open}(J)$ the mean cost of all claims still open, $f_{Mack}(t)$ the Mack factor for the J-the year of development and $Prop_{open}(J)$ is the proportion of claims still open. Table 1 shows the different order of magnitude.

Year	2015	2016	2017	2018	2019	2020
J	6	5	4	3	2	1
$f_{Mack}(J)$	15%	20%	17%	30%	50%	> 300 %
Nb of claims	100	1000	900	1400	300	< 10
$Prop_{open}(J)$	40%	45%	50%	75%	90%	0%
$Dev^{open}_{factor}(J)$	40%	20%	20%	25%	45%	-
$CM^{open}(J) \times Dev^{open}_{factor}(J)$	60 k	65 k	40 k	30 k	25 k	

Table 1: Example of development factor used. All the numbers are anonymized but the authors kept the order of magnitude.

Additional information. For reserving purposes, the insurer provides the number of claims declarations reported at the end of the 1rst year of development, aggregated at the municipality level. One should keep in mind that this type of declaration is independent of the CatNat declaration, *i.e.* even if an important number of claims is declared in the municipality it does not always lead to a CatNat declaration.

2.2. *Geolocation of the building and data at the building scale*

Historically, the insurers' portfolio is not geolocated during the underwriting process. To add new information on the building, the address is used to link the building and the contract. The data provider links the address given with an address geolocated and then associates a building to it. This process is not perfect: the geolocation rate is lower in rural areas and the south of France and uncountable different settings exist. Figure 1 represents the two most common. Even if these last issues are solved, only 80.3% of the addresses are linked to a building. Indeed, around 78 % addresses not geolocated do not have any street number, other addresses are not well reported, not updated, with the wrong spelling, and so on... Finally, out of all the geolocated buildings, an integrity filtering process is done to suppress the wrong building's geolocation or demarcation thanks to different indicators such as the geolocation quality indexes, the number of floors, the footprint surface ... In conclusion, 77.3 % of the addresses/contracts are kept for the modelling.

Figure 1: Copyright OpenStreetMap (Taken the 07/03/2022) - A random street without any relation with the insurer's portfolio. Several issues can be seen. Stars correspond to address geolocation and squares to the related building. Both buildings are not well reported in the building database, with no number being linked to them. Building 1. has an annex and can be easily linked to the 19 Wolfloch Weg Colmar or 19 rue de WolflochWeg Colmar. Tougher, the residential house 2. linked to 23 bis Wolfloch Weg Colmar is 20 meters from the street and is composed of several buildings.

Year	2015	2016	2017	2018	2019	2020
Proportion of claims kept	75 %	74 %	70%	90 %	90 %	0%
Geolocation rate in the area	80 %	75 %	70 %	94 %	80 %	90 %

Table 2: Comparison between the geolocation rate in the department impacted and the proportion of claims kept. The number of claims in 2015 and 2020 is low.

Several checks are done to verify that this process does not bias the risk modelized. Different works on non-climatic risks showed that a bias between 5 and 10 % appears for variables related to the rural areas and is significant. Here, two same conditional frequency models (See Section 3) using only the underwriting variables are fitted, one on the complete database and the other one on the filtered database. No significant changes are seen. In our portfolio, the proportion of claims kept is the same as the addresses geolocated. Because the subsidence risk is a spatial risk, the analysis must also be done by year and spatially. Table 2 shows that lost buildings are not spread uniformly. The geolocation rate is lower in the south and rural area, where the 2015 to 2017 droughts were the most important. On the contrary, the north of France is better geolocated. On the severity side, the risk, by year, does not significantly change. All the areas impacted appear in the dataset used for modelling. In conclusion, the geolocation rate does not significantly bias the risk to modelize.

The data provider hands data on each building and its surrounding; different variables, needing at least the address location, were tested :

Building : Number of floors, living surface, footprint surface, annex's presence, type of roof, altitude, construction's period, energy diagnostics, solar panels ...

Urbanization : Number of building within a radius of 50 meters, the house value, the distance from the closest residential building, the parcel's surface, the vegetated parcel's surface, number of attached houses ...

Location : Type of soil, vulnerability to subsidence risks, distance to the nearest water point, altitude's difference with the nearest water point ...

All variables are coming from French Open Data Sources such as the IGN [1] pictures' database, or ADEME [2] or the BRGM[3] to mention just a few. For each observation, quality evaluations are done and summed up in quality indexes. For missing information or incoherence, the data provider imputes or corrects the values, *e.g.* houses' value is calculated from

[1]The reference public operator for geographic and forest information in France
[2]The French Agency for Ecological Transition
[3]France's reference public institution for Earth Science applications

historical data using its characteristics, the neighborhood's sales... More macro-information at the municipality level was tested, but they are not relevant for the clay shrinkage's risk.

2.3. Gaspar, ONRN and CCR database:

Three external bases are used in this paper; the GASPAR database and information from ONRN about the number of vulnerable houses by municipalities are exploited in the CatNat modelling. Also, CCR historical information helps to compare our results.

GASPAR database. The GASPAR [4] database reports all the administrative procedures related to natural risks: regarding our paper, the declaration of subsidence CatNat risk and the PPRN (Municipal prevention plan for natural risks [5]). This database is of good quality [6] for most of the year. Some corrections are done for older years for which PPRNs were wrongly reported. If before 2000 the notion of subsidence CatNat was not clear enough, the database's quality is very clean starting from 2005. The CatNat information is available through several indicators, *e.g.* the starting date of the disaster, the duration of the episode, the end date, and the date for which the disaster is declared as a natural catastrophe. Looking through the different years, the evolution of CATNAT's declarations is evident. Before 2003, a subsidence event may exceed a year and the mean declaration delay was a lot higher than nowadays. The stationarity starts between 2005 and 2009 for different indicators, as exemplified by the duration (Figure A.17).

ONRN database. The ONRN [7], the French National Observatory of Natural Risk, made available for each municipality the number of building subject to a subsidence's risk. The information corresponds to the 2015 period and to the 2015 Subsidence vulnerability map. At that time, only three class exist "*none*", "*low*" and "*high*". 3 (+ 3) variables are created from it, counting the number of houses by risk's class (resp. proportion). The downside is that the vulnerability map is not the last one. Our model will also use similar variables corresponding to the number of insured houses in the portfolio in each class of risk (from the last vulnerability map class "*none*", "*low*", "*medium*" and "*high*").

The CCR database information . The reinsurer CCR provides aggregated information by municipalities from 1995 to 2018. The ratio claims amount and premium, the mean claims cost, and

[4] *fr: gestion assistée des procédures administratives relatives aux risques naturels*

[5] *fr: Plan de prévention des risques naturels*

[6] Few quality tests were done on subsidence risk whereas for floods risks the database has been tested in broad terms *e.g.* [Casaux et al. 2019] [4] or for a spatial-temporal quality problem on the side of the flood by [Douvinet and Vinet 2015] [6]. Nonetheless, the latter issues are similar to drought ones.

[7] *fr: Observatoire National des Risques Naturels*

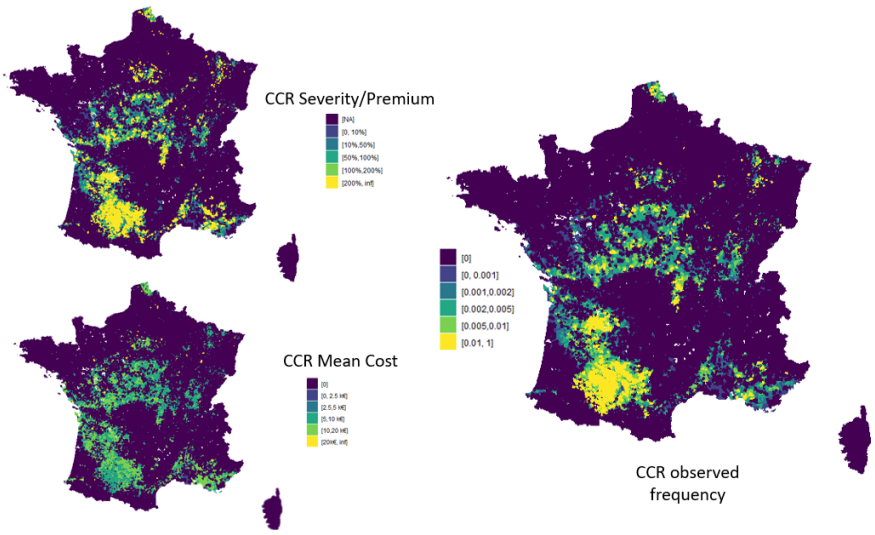


Figure 2: Information provided by the CCR on the subsidence risk on the French market. The data correspond to aggregation from 1995 to 2018 sinistrality.

frequency are available. In this work, they are used to compare our results, highlighting our models' limits. Unlike our data, the claims are coming from different insurers all over France, and consequently, it can be considered as market information. The three maps - Figure 2 sum up all the information provided by the CCR. This information will be used to test our models and not as variables.

2.4. Drought indexes and meteorological variables

Around 30 meteorological indicators were available on the municipality scale from 2010 to 2018. Each variable is available annually and a variable referring to the mean value out 2010-2018 historical is created. The meteorological information was unknown for 2019 and 2020 and was replaced by the mean value calculated on the 2010-2018 historical data.

Non-exhaustive list of meteorological data used (NbD = Number of day for which);

Temperature (T°): NbD $T° \leq 18$, NbD $T° \geq 34$, NbD $T° \geq 4$ + mean seasonal temperature, NbD $T° \geq 8$ + mean seasonal temperature, ...

Precipitation (Prep): NbD *Prep* ≥ 1 mm, NbD *Prep* ≥ 1.25 mean precipitation, NbD of heavy precipitation, prep quantity...

Frost : NbD of superficial frost, NbD of deep frost, NbD consecutive of deep frost, NbD with snow ...

Wind : NbD of heavy wind hust, Number of wind gust with an hourly speedy higher than 99 $km.h^{-1}$, ...

Note that not all the meteorological variables were relevant and some indicators are highly correlated.

Besides meteorological variables, drought indexes were also added by year at the municipality scale. The drought indexes' literature has proposed several indexes to understand the connection between drought and weather. Different types of drought have been proposed by Wilhite and Glantz (1985) [19]: agricultural, meteorological, hydrological, and socio-economical drought. The operative event of clay shrinkage is mostly brought by hydrological and meteorological drought. A subsidence event is mostly triggered by a clay shrinkage and then a clay's swelling is provoked by higher humidity such as rain. The quick level shift cracks the foundation or the walls. These damages are covered in the household insurance contract if the municipality has declared a natural catastrophe. Recently, a clay shrinkage-swelling natural disaster is declared if the meteorological indexes are abnormal, corresponding to a period of return higher than 25 years. Since 2010, the indicator SSWI is used at the *inter-ministériel* meeting to accept the status of natural disasters. To evaluate the probability that a municipality declares a CatNat, three indicators will be studied. All these indexes were available monthly and extrapolated to each municipality from 2010 to 2018. All this information has been summed up into three variables (*severity*, *duration*, *magnitude*) by calendar year and by drought index.

SPI : Standardized precipitation index is commonly used, *e.g.*[18]. ([14]) has developed a methodology for standardized indicators. In our case, we consider the **SPI - 1 month** the annual monthly minimum named *magnitude* (Similar ESPI variable from [5]), maximum duration of events, *duration*, when SPI is below -1 and the mean of the SPI during the event named *severity*.

SSWI : The standardized Soil Wetness Index (SSWI) is better adapted to agricultural drought. The basis SWI is used as a meteorological criterion from the CatNat declaration since 2009. We consider the **SSWI - 1 month** the annual monthly minimum, the maximum duration of events when SSWI is below -1 and the mean of the SSWI during the event.

RDI : Reclamation Drought Index was developed by the United States Bureau of Reclamation in 1996 to trigger drought emergency relief funds associated with public lands. This index captures drought severity as well as the duration and can be used to predict the start and the end of drought periods. RDI uses temperature and hydrological components, incorporating evaporation into the index for its calculation. We consider the **RDI**: the annual monthly minimum, the maximum duration of events when RDI is below 0 and the mean of the RDI during the event.

Annual but not seasonal : Contrary to [5], our goal is to predict the frequency and the claims for an insurance company that consolidates its financial statements in late December. For reserving, the calculus is done for each year. Moreover, the different models used suppose independence of each observation, which can not be assumed seasonally. From the Gaspar database, nearly no municipality had two CatNat declarations in the same year. This is mainly explained by the delay of CatNat declaration. The period event could be increased to take into both events and all claims incurred are declared in the first event. Two limits of this method must be stated about the

- Inclusion of droughts starting at the end of the year and ending at the start of this new year: few in number but not insignificant;
- Meteorological criterion updated in 2018 has now defined a new seasonal threshold. Even if we used the standardized index which partially deals with this problem, our model is based on historical events before 2018.

The idea is to use these three variables, to sum up, the better possible, the worst drought of the year according to each indicator. The downside of these three variables is their dependencies, which are not trivial and impact the way the linear model should take them into account.

As explained by [5], the subsidence meteorological criteria changed through time. From 2000 to 2003, only a hydrological criterion was used. The year 2003, the worst year for subsidence, was not captured by this criteria and since 2004 a meteorological criterion was used. Because the criteria were not easy to apprehend, the SSWI was finally used starting in 2009 evaluated in Winter, Spring, and Summer. In 2018 a new threshold is used for each season. It is important to put forward that the justification is quite succinct, and the clay presence and the number of houses damaged may be better key drivers than this meteorological factor.

2.5. Particularity of the portfolio studied

The contract portfolio used covers the entire mainland France from 2015 to 2020. With an exposure superior to 700 000 per year, the claims are spread out in all the potential territories impacted by the subsidence risks according to the vulnerability map provided by the BRGM. According to the insurers' actuaries, the claims frequency of 2016 is unusually high compared to other insurers for the same year. The CCR insurer has also pointed out this particularity. A probable underlying reason is that several buildings were damaged before 2016, but waited for the CatNat declaration in 2016 to be declared. The mean contract's seniority is about the same as the French market one, around 10 years with a constant retention rate.

The data have undergone some temporal evolution; the underwriting process revised through the years. For instance, period of construction variables has new modalities, leading

to lower completeness. Evolutions due to the underwriting process are not exceptional. Hence, some variables may have a limited impact, *e.g.* the insurer's period of construction will have a lower impact than the one gathered from open data with completeness around 97%. Before 2015, the number of rooms and the house's surface were asked. However, for the new underwriting process, only the question about the number of rooms is still asked. Therefore, missing values on the surface are imputed using the number of rooms and also information from the geolocation process.

Finally, the management of claims also impacts the data. For each contract, an estimation of reconstruction value is calculated using the underwriting information. If important damages happen such as subsidence damages, an insurance expert evaluates the damage and also gives a proper evaluation of reconstruction value. In this process, the surface is sometimes set to 0. Therefore, the authors were not able to train proper Random Forest, XGBoost on the conditional frequency models as explained in Subsection 3.3. Using external information, the surface gathered during the underwriting has been correctly imputed to be used in the GLMs. The authors would like to highlight that the linearity of GLMs allows avoiding learning this causal information and also helps to evaluate the bias due to our model for surface imputation.

3. Modelling the CatNat frequency, claim cost and legal declaration

This paper's modelling process is based on the CatNat regime process. The indemnity occurs only if the municipality has declared subsidence natural disasters. Hereafter, within the same municipality, the frequency and the claim cost depend on the building characteristics. The first step is to modelize the municipalities' CatNat declaration in Subsection 3.1. After a filtering process to clean up the errors on the link between building and geocoding (Subsection 3.2), models for claims' frequency (Subsection 3.3) and claims' cost (Subsection 3.4) are done with and without the geolocation of the building. From a performance perspective, the improvement thanks to the data is significant and the aggregation of the different models leads to several insights. All the models are created to be used operationally. Machine Learning models such as Random Forest ([3]) or XGBoost ([8]) are used with frugality for reserving and are almost forbidden in underwriting use. Therefore, the combined model are using GLMs with a limited amount of splines. For models on the CatNat declaration, we used a XGBoost-based model, which can be justified due to the complexity of meteorological variables and the use of compositional variables for urbanization impact. To be used in an underwriting process or operational uses, the results are resumed in zones similar to a zoning variable added in premium models. Figure 3 sums up the results.

3.1. Natural Catastrophe (CatNat) declaration models

CatNat models focus annually on the CatNat declaration at the municipality level. For reserving purposes, the probability that the municipality declared a CatNat knowing annual

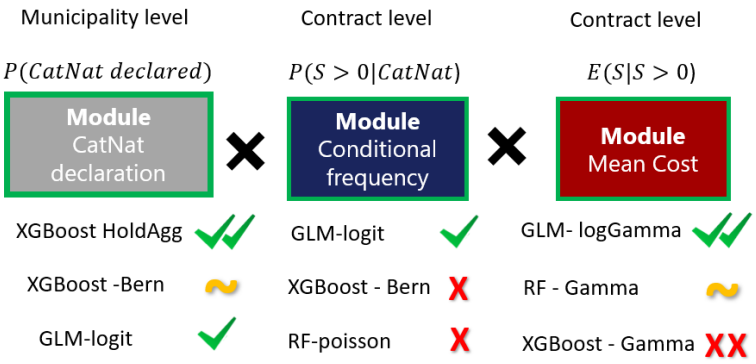


Figure 3: S represents the cost of a claim, and *Ber* is an abbreviation for the Bernoulli distribution. Due to data management, XGBoost and RF are not suitable for conditional frequency, even the GLM must be used carefully. Moreover, the mean cost modelling is impacted by the reserving process resulting in improper convergence for the RF and even worse for XGBoost. For CatNat declaration, GLM and XGBoost have proper results. Nonetheless, the latter may learn improper causality. Thus, we propose a model less performant XGBoost HoldAgg but more robust.

meteorological indexes at the date t and the date $t-1$, $\mathbb{P}(\ CatNat(mun,t)|\ meteo(mun,t,t-1), charac(mun))$ is modelized.The same probability without any meteorological information $\mathbb{P}(CatNat(mun,t))$ is calculated with historical information available, as follows :

$$\mathbb{P}(CatNat(mun,t)) = \sum_{t=2010}^{2018} \mathbb{P}(CatNat(mun,t)|\ meteo(mun,t,t-1), charac(mun)) \times \mathbb{P}(meteo(mun,t,t-1)), \tag{2}$$

where $\sum_{t=2010}^{2018} \mathbb{P}(meteo(mun,t,t-1)) = 1$, mun a municipality, t a year, $meteo(mun,t)$ the weather indicators, $charac(mun)$ the other characteristics of the municipality and $CatNat(mun,t)$ if there is CatNat declaration in the municipality at the date t. This Equation (2) assumes that all the drought scenarios possible appear between 2010 and 2018 and conditionally to the municipality information (urbanization, meteorological index, ...), and that there is no spatial dependency [8]. Let's assume that all probabilities are equal, one with better knowledge could adjust the different scenario's weights.

The training method. To find the hyperparameters, a spatial k-fold approach for each model is done, where the model is fitted on 50 % (70 % for GLM) of all regions and validated on

[8] i.e. having a municipality nearby declaring a subsidence CatNat does not increase the probability to declare a CatNat all other things being equal.

the regions left out [9]. The approach of time cross-validation ([1]) performed by removing the future from the analysis done was not relevant to our data. Starting from 2010, until 2016 only the year 2011 is a drought year. Moreover, the variables considered in this paper do not all have the same spatial properties. Indeed, urban or meteorological variables are spatially correlated. Therefore, using a similar spatial and temporal model for GLM (as in [5]), some coefficients were volatile and not interpretable, *e.g.* the coefficient of SPI severity, magnitude, or the number of building in a radius of 50 meters.

Reserving models used. First, the CatNat declaration models were fitted using the number of claims declarations, recorded the same year at the end of December and aggregated at the department scale. In this model, the annual meteorological variables are not used, only the historical mean and drought indexes. For the geolocated variables, aggregated information at the municipality scale is used; the relevant ones are the mean altitude of all the portfolio houses, the mean number of building in a 50 meters radius, the mean number of houses highly vulnerable to clay shrinkage (resp medium, low, none), the mean distance to watercourses and the mean probability of being the main residence. The models' results can be shown in Figure 4 from 2001 to 2020 using an XGBoost model.

Figure 4 shows that the model is correct. However, the subsidence declaration does not always lead to a CatNat declaration. Plus, some clients wait for the CatNat declaration to declare to their insurer their damage. Consequently, if the number of CatNat declaration at the end of December is informative, the structural dependency does limit the model's performance. For instance, for 2004, 2010, and 2013, few CatNat declarations have been enforced due to the drought period return. Nonetheless, several claims damaged were declared inducing an overestimated probability. By adding all the drought indexes, the variable - the number of declarations in the first years of development - was no more relevant, all the information being captured by other variables.

To improve the model, the idea is to add meteorological and annual information. Beforehand, two steps are done for the variables' selection. To keep influential variables, a simple XGBoost and Random Forest are fitted. Afterward, a stepwise logit-GLM (Efroymson, 1960 [7]) selects all the non-meteorological variables. Finally, an XGBoost is fitted using the selected variables, the three indexes and two meteorological information (at the most - here the number of days for which the temperature is below 18 degrees and the annual precipitation quantity.).

[9] The spatial grid used is the *department*[10], other independent grids were tried (set by hierarchical cluster analysis (HCA)) with 70, 500, 1170 groups. The changes are not significant enough (same hyperparameters/same likelihood).

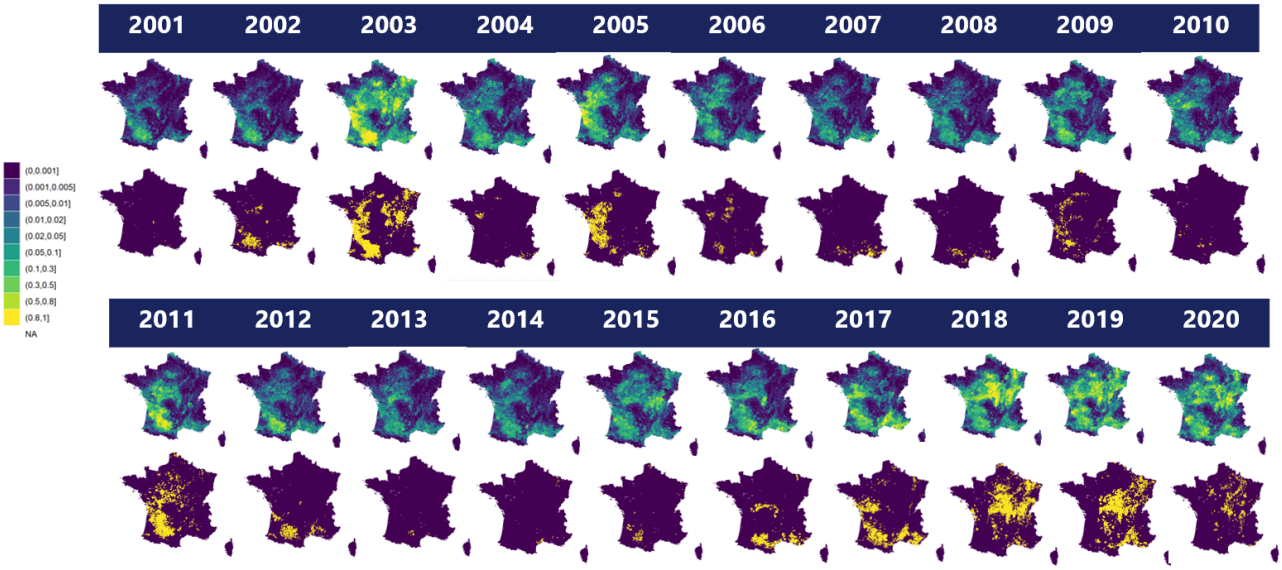


Figure 4: CatNat declaration modelling at the municipality scale for reserving purposes. The second row corresponds to the observed CatNat declaration. This model does neither use annual drought indexes nor meteorological variables but the number of subsidence damage declared at the end of December at the department scale.

To be used in reserving or for pricing purposes, the model must be fully interpreted, especially when using variables changing each year. If the linearity of GLMs helps to fully understand quickly the learned structure, XGBoost's created structure is not easy to apprehend. To partially understand the model, Shapley values [11] are used to find the cross-effect for each model for meteorological interactions and compositional/aggregated variables. Figure 6 shows interactions between the 4 most important variables according to Shapley score or the traditional importance plot based on the Gini importance. Figures in the Appendix C also show that the previous meteorological variables add some relevant information and that the Shapley interactions values can difficulty consider all the tri-variables (or more) interactions learned.

Table 5 shows through the GLM linearity that drought indicators do pretty well to determine the location and the frequency on their own. Thus, the performance of the XGBoost may be explained by the use of drought indicators. In our case, the authors trained several

[11] package *SHAPforxgboost*.

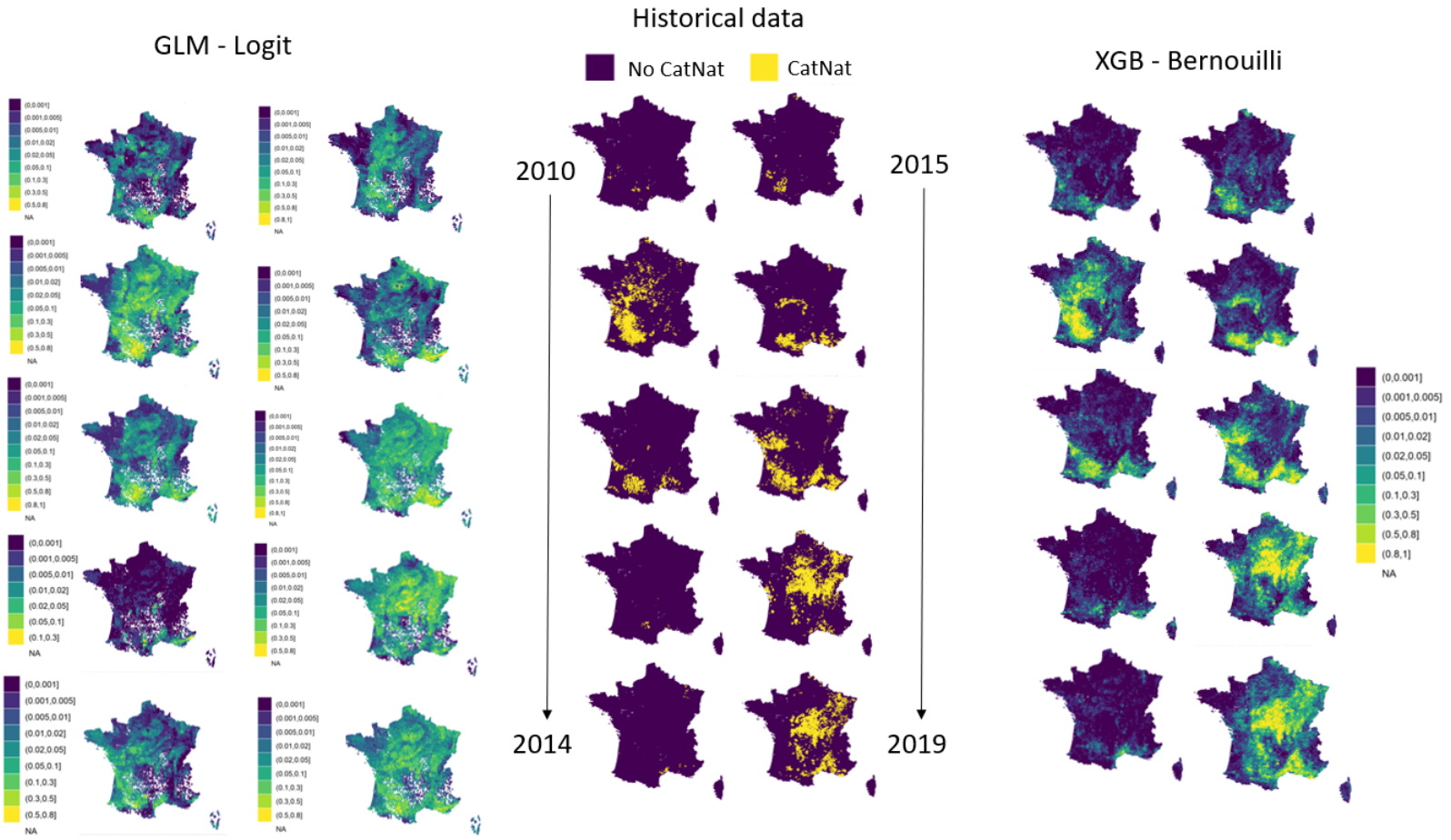


Figure 5: Comparison between the observed CatNat declaration, the GLM logit (using RDI magnitude, SSWI severity and SPI magnitude) and the XGBoost using all the variables calculated from the three indexes.

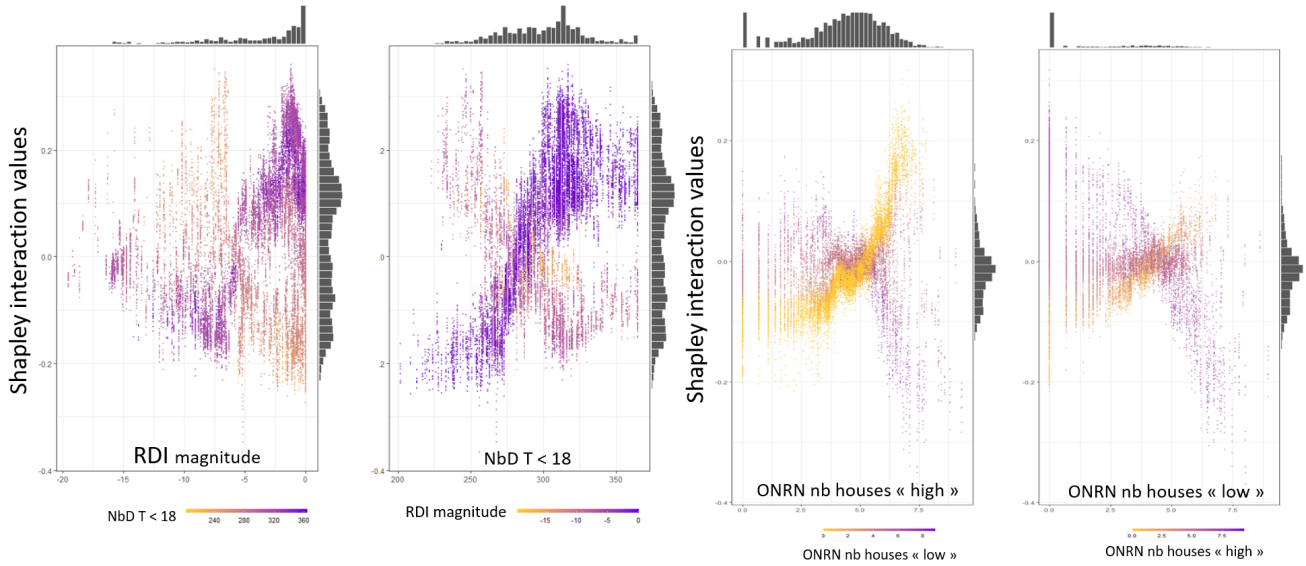


Figure 6: Shapley's interaction between meteorological variables and aggregated variables using a simple XGBoost trained on 2010-2018 historical (20 000 rows for the calculus of Shapley values). The ONRN variables are represented in the log scale. The analysis is only bivariate; the interaction between all the meteorological variables is impossible to completely discerned. A high number of interactions are present in the XGBoost learned structure and are difficult to replicate in a GLM or a GAM.

XGBoost models, one on 2010-2018 historical data, a second on 2010-2019 historical data, and a last on 2010-2020 historical data. It is important to remind that the meteorological information is unknown for the years 2019 and 2020 and was replaced by the mean value calculated on the 2010-2018 historical data. The model should be less performant if the 2019 and 2020 years are added. Table 5 shows that it is not the case; this is a very problematic issue and shows that the XGBoost, or is not predictive, or has learned non-causal information. Therefore, the author added a constraint to use a model.

Operational constraint *For the years 2020 and 2019, the performance of a model should be lowered or equal to the GLM performant trained from 2010 to 2020.*

Inspired by Maillard et al. 2021 [13], the next structure proposed verifies empirically our operational constraint. The method "HoldAgg" consists of cross-validations and averaging. Several XGBoosts are learned. Each XGBoost is trained on 6 years and the hyperparameters used are the ones maximizing the cost metrics calculated on the 2 years left aside. The final result considered is the average value of each XGBoosts' results. One can remark that the mean value has no reason to be a probability. Therefore, we learned a classifier (Niculescu and Caruana, 2005 [15]) mapping the average of the result to a probability. The classifier is

learned on the 10 % of the training set on the mean value by step of 10^{-3}. Therefore, the averaging model is less performant than the simple XGBoost one. Yet, fitting this XGBoost HoldAgg model on the years between 2010 and 2020, the performances for the years 2019 and 2020 are lower than GLM ones.

This method uses the "mrl" package [2] with Gaussian Process to find each hyperparameter more efficiently. First, 30 hyperparameters settings are fitted. Then, a Gaussian Process is fitted on all 30 points to find the next point which leads to the best-expected improvement. An XGBoost is calculated with this hyperparameter setting. Then a new Gaussian Process is fitted on the 31 points and so on ... We repeated the process 30 times.

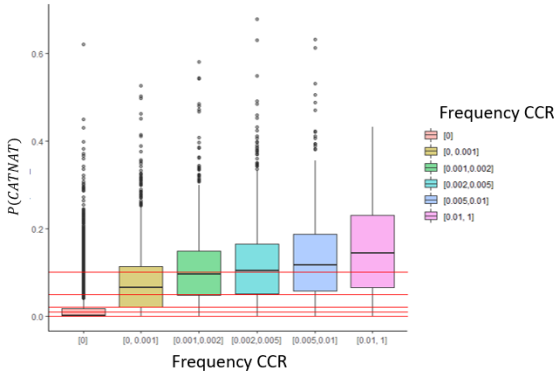
The difference with the aggregated Hold out :. Appearing at first in NeuroImaging ([9]), the Aggregated Holdout process, or more simply "CrossValidation + averaging" has been proved theoretically under several assumptions by [13]. Performance of each XGBoost depend on the training and testing dataset, *e.g.* a XGBoost trained on all years except 2013-2014 (two years without any drought) the choice of the hyperparameters leads to the least performant model. Our model does not verify all the assumptions needed, which are used to justify the theoretical performance of the Aggregated Holdout process. The assumptions needed are :

- Classification with the 0–1 risk or convex risk: Our variable is a binary one, but we used a regression model to calculate a probability.
- The split for the hold-out cross-validation is temporal and is correlated with the variable to modelize. Indeed, the number of CatNat declarations depends on if the year is dry or not.

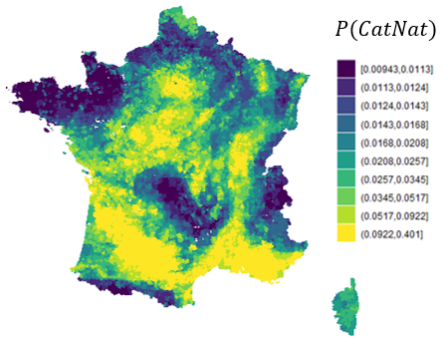
Model results. Comparing the XGBoost results by year with the GLM results using all the non-meteorological variables and the RDI magnitude variable, one can see all the years are well-considered. There are several limits. The year 2018 is not perfectly captured by the meteorological indicators, and for years without drought, a residual probability still appears.

Having a probability, we have re-simulated the number of natural disasters declaration and compared it to reality. Low probabilities ($< 5 \times 10^{-3}$) are overestimated; this comes from the classifier's precision. For the higher probability, the model is well-calibrated.

Moreover, the results can be compared to the CCR historical return period on Figure 7a. The model matches the CCR information, where mean values increase with the period of return given by the CCR. The means are not in the interval period of return because the historical period considered is different.



(a) Comparison of the aggregate value and the historical CCR frequency. Remark that our model is only based on the 10 most recent years and the CCR on 23 years. The horizontal lines refer to the CCR ranges' limit.



(b) Estimated probability from the XGBoost HoldAgg. 10% of French municipality has a probability to declare a clay shrinkage CatNat superior to 0.1.

Figure 7: The last historical years represent pretty well all the drought scenarios possible. Our assumption is that each year/scenario has the same probability to reappear, meaning that year with drought appears 3 out of 8 years. Looking to the 2001-2020 years, the probability would roughly be around 7 out of 20.

3.2. Integrity filter

Before working at the address level, it is important to analyze the errors in the data. Indeed, the geocoding's accuracy is not perfect due to the geolocation process, the address's quality, and the databases used. Therefore, an integrity step is done before modelling. Here, the integrity rule is : "All the building geolocated are residential individual houses". Using the number of floors, the data provider's information, and the potential surface, 3.4 % is set aside. Quality tests on geolocation are done on the houses geolocated and on the addresses not geolocated. For addresses not geolocated, the results are summed up in Figure 8. Using the same process for geolocated buildings, we can state that 9 % of the addresses are potentially not well-geocoded and within these less than 4 % of the addresses are not linked to the "good" buildings in our judgment. Unfortunately, no direct rule can separate these addresses. Therefore, these addresses are kept, lowering the marginal impacts of variables using geolocation[12].

Several analyses are done to verify that filters do not impact significantly the sinistrality. The number of claims lost is about 20% such as the proportion of not geocoded buildings. The claim distribution is not impacted by the claims lost. The lost buildings are mostly linked to rural and south areas, for which the standardization of addresses has not been done yet.

[12]Not all variables are affected in the same way.

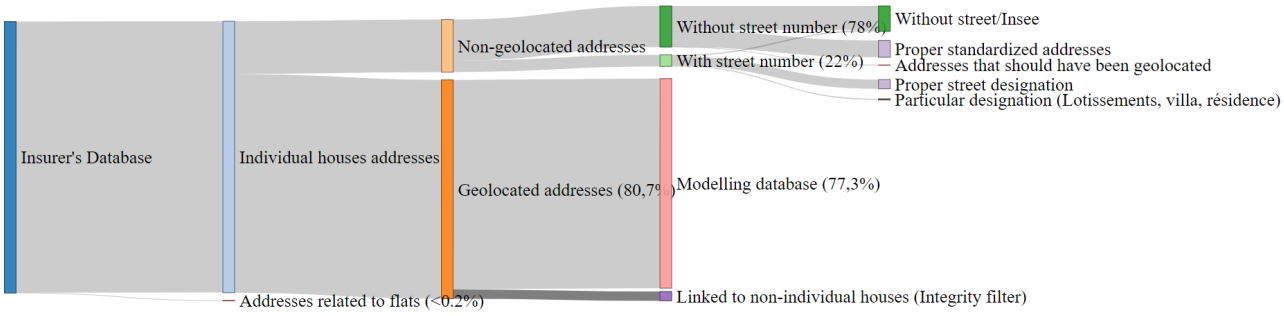


Figure 8: Controls on the geolocation process for non-geolocated addresses. A similar process is done for geolocated addresses; a stratified sample is chosen to compare our geolocation and the one Google geolocation would give.

Consequence on the modelling : To control the filtering process, two models using the insurer's variables are done: one on the complete database and one on the filtered and geolocated database. No coefficients' changes are significant. In conclusion, the geocoding process does not influence significantly the subsidence risk according to the insurer's information[13]. For contracts not geolocated, the conditional frequency and claims will be calculated using the department mean value.

3.3. Conditionnal frequency at the building scale

In this Section, the goal is to model the frequency of houses conditionally to a CatNat declaration $\mathbb{E}(Card\{S > 0\}|CatNat\ declared)$. To compare the performance gained from geolocating contracts, two models are done, one with only underwriting variables named *Insurer model* or $Referent\ model$ and the other one using all variables available named $Performant\ model$[14].

Distribution used : The maximum annual number of claims in the database is equal to one. Therefore, we used a Bernoulli distribution in our XGBoost and logit-GLM. For the variable selection, a Random forest using Poisson distribution is used from the *distRpackage* because of the credibility hyperparameters options [15].

[13]This statement could be discussed according to the data coming from the geolocation of buildings. However, by definition, the filtered buildings are wrong so nothing more can be done.

[14]This model is named "performant" because by construction the model is more performant than the $Referent\ model$.

[15]For low probabilities, Poisson and Logit are good approximations of each other, see for a robust modified Poisson model for binary data [20]. The credibility hyperparameters of RF Poisson proposed, leads to good results.

Surface and the reconstruction value variables of the insurer : The data analysis has pointed out that a surface null or a precise reconstruction value were highly linked to subsidence claims. The expert during the claim process can modify the data sent, adjusting the reconstruction value and suppressing the surface value [16]. Therefore, the null surface value has been replaced by a prediction using an insurer's variable (number of floors) and several other geolocated variables (living surface, construction period, ...). The imputation is not perfect, especially for extreme values. For GLM, it is important to have in mind that it underestimates the frequency claim for building with a low surface and very important surface. However, for machine learning methods (XGBoost or RF), the use of the surface and reconstruction value variables is very problematic. Looking through univariate and bivariate partial plots, it seems that these types of models are learning non-linear relations and worse, are learning not operationally justifiable structure, *e.g.* reconstruction values between 155k and 165k have an important marginal impact on subsidence.

The variables used for the performant model are presented in Figure 9. The *Referent model* uses the same underwriting variables and also construction's period, the value of the personal property insured, and the number of rooms that were not very relevant and hard to justify.

Figure 9 represents the deviance gained when adding the variable to a GLM model, including all the others. Only three insurer variables are still relevant. Being an owner non-occupant highly reduces the subsidence declaration. Indeed, the insured person is less careful than an owner occupant. The age also impacts the declaration, older people declare also less, but the marginal could also trigger by correlation with the type or the period of construction of the house. The surface imputed is not truly an insurer variable, the missing values being imputed by different variables needing the geolocated building especially period of construction, surface of the building, and the footprint surface which are in the performant model. Remark that the insurer variable can not be totally replaced. Indeed, it corresponds to all the insured surface (not only the main building related to the address) with the annexes and refers also to the "used" living surface. Plus, no geolocation errors exist.

Results Table 3 shows that by adding information thanks to the building geolocation, the performance for several metrics has improved. Even by adding a zoning variable through the credibility spatial smoothing methods, the *Insurer model*'s model is still less performant than the *Performant model*. Moreover, no significant zoning variables could be added to the *Performant model*. All the geographical information is captured at the address level by all the other variables. The geographical information captured can be seen by looking at the risks

[16] In fact, the building surface stopped being asked a few years ago. Only the number of rooms is still being asked. The reconstruction value is a created variable using the surface or the number of rooms and other geographical information.

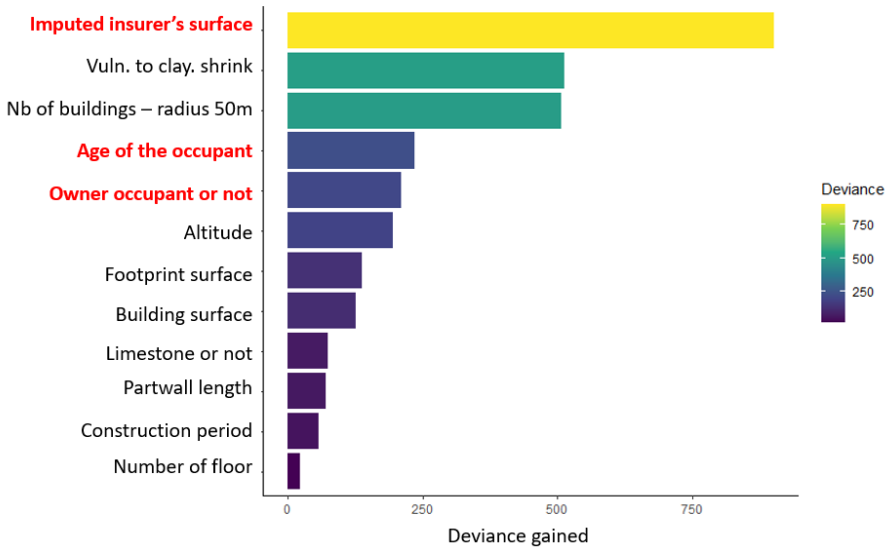


Figure 9: Performant model : Deviance gained when adding a variable to all the other ones in the GLM framework. One should keep in mind the dependencies between the variables. All variables are validated using a type III test, univariate graphs like the ones in the Appendix B, and have a causal explanation.

maps - Figure 10. An important spread gap at the municipality level is observed. Remark that the performant model also segments within each municipality and not the Insurer model.

The geolocated variables used are very segmenting, *e.g.* the more floors number, the deeper the house's foundation is and the less subject to subsidence risk the building is. The vegetated surface or the number of building in a 100 m radius also captures the tarring level of the neighborhood and the presence of canopy probability. The vulnerability to clay shrinkage is also very segmenting (see appendixes for some univariate plots)[17].

Limits The principal limit of this conditional frequency model comes from the data. Indeed, for recent years, the claims are not all declared. Therefore, the frequency is slightly underestimated. Moreover, no meteorological information is relevant for the conditional frequency, perhaps the information is already captured in the CatNat declaration module. It might underestimate the frequency for drought years and overestimate for years without extreme temperature.

[17]The precision is in-between the street and the iris level. Indeed, the BRGM map is not précised enough by construction.

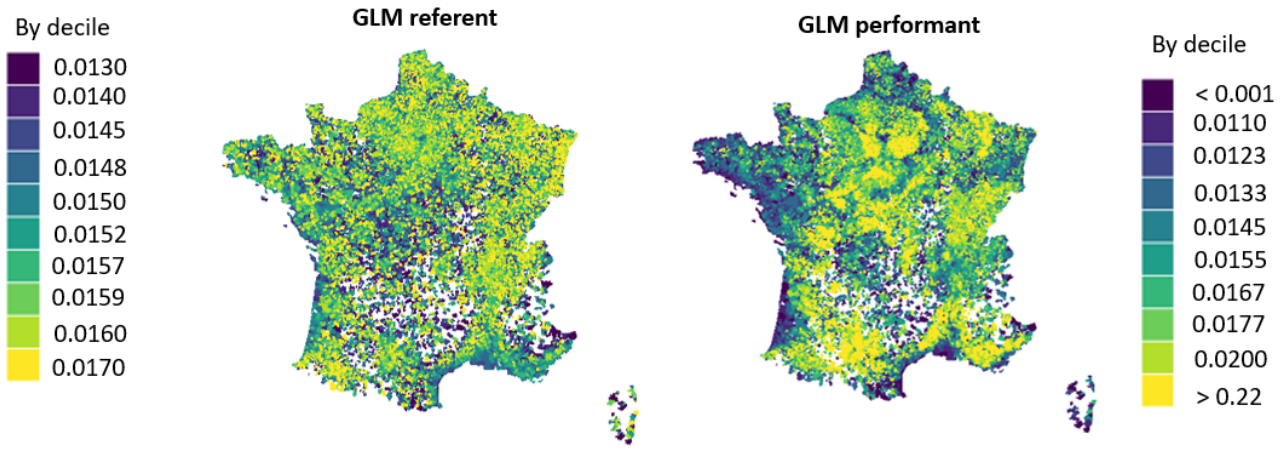


Figure 10: Maps about the predicted conditioned frequency (if there is a CatNat, the probability of a house to be damaged). The colours are determined by deciles of the mean value of each house in the municipality, and the values correspond to the mean in the category. Probabilities are less relevant in zones where drought CatNat are inexistant such as brittany.

Model	Type		EDR[a] Poisson	EDR Logistic	R^2	Gini
GLM - binomial logit	Referent	Without insurer's surface	100%	100%	100%	100%
	Referent	With the imputed surface	+ 42 %	+ 71%	+ 133%	+ 22%
	Performant	Without insurer's surface	+ 130%	+ 91 %	+ 379 %	+ 39%
	Performant	With the imputed surface	+ 164%	+133%	+ 430%	+ 48%

Table 3: Comparison between the metrics for the conditional frequency models. Models optimize logistics deviance or equivalently the EDR logistic. The R^2, the normalized Gini and the EDR Poisson are used to control models.

[a]The EDR is one minus the ratio of deviance between the model one and the satured one.

3.4. Cost models

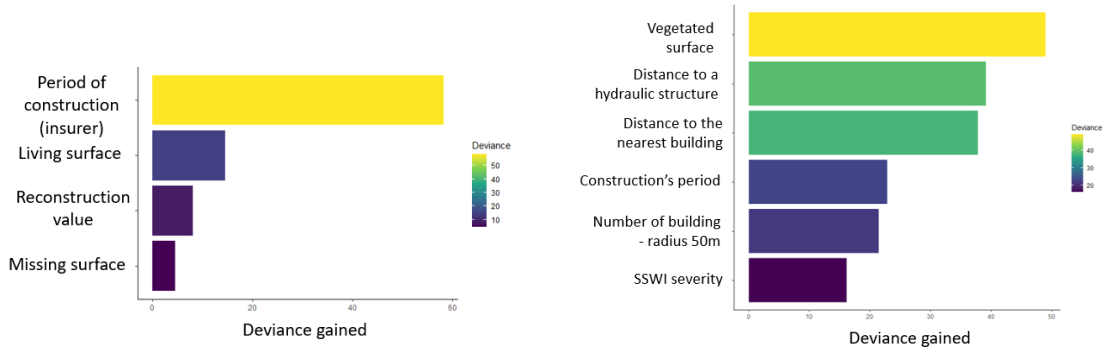
As for the frequency model, two types of model are considered: *Referent model* (Insurer's model) and *Performant model*. As explained in Subsection 2.1, most of the claims costs are yet open. Therefore, several databases and methods were tested (using the year as a modality, looking at a ratio per year ...). Finally, the projection of costs still open gives the best results, albeit the different linear impacts are very similar for each method. All these controls help to validate the database's robustness. One claim was set aside due to the asbestos presence, leading to an extreme cost.

The claim database used for modelling contains only claims which had payment or for which an expert has evaluated the claim cost. For the others, an automatic opening claim is provisioned, waiting for an insurer's expert to estimate the subsidence damage. These claims are not informative and thus, not used in mean cost modelling.

XGBoost models do not converge to the mean value but more to the median one due to the claim cost particularity. Therefore, GLM and Random forest Gamma will be used. For nearly all metrics — Table 4, the *Performant model* is better than the *Referent model*. Using one-way analysis and GLM linear structure, margin impacts of added variables are new and relevant, replacing the insurer variables (See Figure 11).

Model	Type	EDR Gamma	R^2 (10^{-2})	Gini	Mean difference
GLM-logGamma	Referent	(0,0052)	(-0,3)	(5,98)	- 418
RF-Gamma	Referent	100 % (0,015)	100 % (1,02)	100 % (9,35)	- 451
GLM-logGamma	Performant	- 20 %	-70%	- 24,3%	-13
RF-Gamma	Performant	+ 46 %	1.59	94,4%	-8

Table 4: The referent models are not very performant. The referent's RF might have learned improper marginal correlation using also the socio-professional information (operationally not justifiable). The referent's GLM has an artificially high normalized Gini linked to his poor performance. The performant model does not use the SSWI severity variable. The latter when used, does not significantly alter the performant models' metrics.



(a) Referent model GLM-Gamma

(b) Performant model GLM-Gamma: The variable SSWI severity is only used for reserving purposes. Other variables are set aside; their marginal impacts not being relevant or justifiable.

Figure 11: Deviance is gained when adding a variable to all the other ones in the GLM framework. One should keep in mind the dependencies between the variables. All variables are validated using a type III test, univariate graphs like the ones available in the appendices, and have a causal explanation.

Limits Three limits can be stated. The first one is that the mean cost is driven by the development factor. Thus, the average mean cost of 27 k€ is volatile with an IC around 3k€. This development factor dev_{factor} covers the fact that open claims are less informative than if there were completely paid. Therefore, the marginal impacts of the variables are very likely underestimated. Finally, the dev_{factor} correction factor is done by year. Therefore, some meteorological information such as the RDI severity cannot be added. Indeed, this latter is too correlated with the annual year. Therefore, the learned marginal impacts are difficult to explain properly. In short, RDI severity and RDI magnitude could potentially be used in cost claims if we had more historical data or better-developed claims. The latter limit impacts the period of construction variable. Indeed, this variable is linked with each year through spatial correlation. Claims occur depending on the year in different areas, adding an artificial dependence to the true marginal impact of the construction's period variable.

These different limits and the operational constraints lead to prefer the GLM models. In short, the non-linear structure of Random Forest hindered the control of the impact of undeveloped claims. However, we will use the RF for the referent model (because the GLM one has a too low performance) and the GLM for the *Performance* one.

3.5. Aggregate the models

Figure 12 shows that the referent GLM is not performant and also shows that the Random Forest models neither for the referent model nor the performant model did converge properly to the mean value. Nonetheless, the latter segments are better than the GLM one. Remark

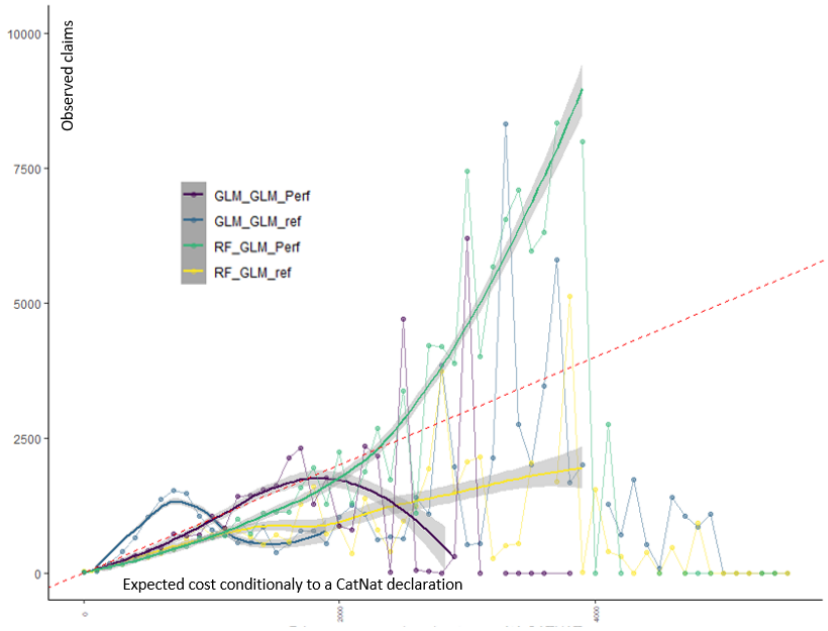


Figure 12: Comparison between the different models combining the conditional frequency and the cost. For each type of model, a linear model using a spline and the exposition of each point is fitted in order to perceive the trend.

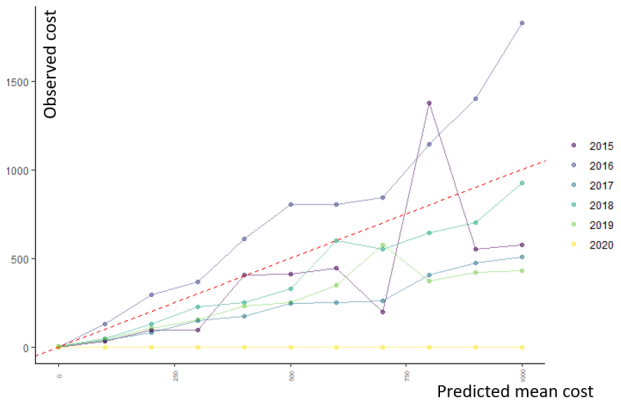


Figure 13: Comparison between the observed sinistrality by year and the performant model (GLM + GLM + XGBoost HoldAgg).

that the highest stable value is around 1 k€ for the referent model (RF + GLM) and for the performant (GLM +GLM) it is around 2,5 k€.

Adding the CatNat modelling for reserving, the same graph 13 can be done by year. One can remark the linearity between the observation and the prediction. All the predicted premiums higher than 1 000 euros are regrouped together. Year 2020 is not yet developed (less than 5 closed claims). The 2016-year historical claims are higher than the one predicted as explained in Section 2.5. The 2015 year is much more volatile, being the year the less exposed to drought. For the expected cost for "pricing" purposes, the CatNat model used will be the same for the *Referent model* and the *Performant model*.

Several limits must be stated about these models.

First, in our model, frequency and claims cost do not depend on previous claims. In our data set, no building has declared twice a subsidence claim. According to the insurer's actuaries, on historical data since 2001, some contracts have declared more than one damage; often the first damage being less costly than the second. Two questions stay unanswered: does the damage correspond to the same building? Does this process correspond to buildings poorly repaired? If both questions are answered positively, the stationarity supposed in our models may not be verified.

Secondly, the conditional frequency model is calculated on data containing 2019 and 2020 from the April-2021 point of view. Some contracts have yet to declare their subsidence damage. Our model may underestimate conditional frequency.

Thirdly, the cost claim model was based on a high number of undeveloped claims, especially in 2019. Therefore, variables having annual variation, such as meteorological indexes, were set aside due to the lack of information.

Then, the modelling of the CatNat declaration depends on the historical range considered. Using all the data at disposable, the range of 2010 - 2018 has been considered. Nonetheless, one could argue that a wider scale could have been used, e.g. starting from 2003 (which is the worst year for subsidence). The stationarity of the CatNat declaration and the robustness are open questions, for which the authors have limited knowledge to answer properly.

Lastly, the legal evolution of the CatNat regime and climate change impact the stationarity of the process. This part will be discussed in the last Subsection 4.3. In short, we would like to emphasize that the models developed have a predicting power over a short period, even if different tests on other insurers' household insurance portfolios have validated the results robustness.

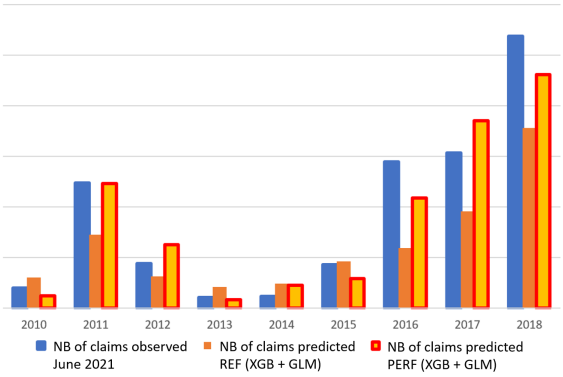
4. Reserving, prevention and insurability statement

In France, the pricing of CatNat coverage is not allowed. The associated premium is calculated as 12% of the sum of the damage coverage premiums. Nonetheless, modelling of the expected cost is compulsory for the reserving process (Subsection 4.1) and can be used to identify the highly exposed segment for instance for prevention purposes (Subsection 4.2). The downside of the model's performance is the insurability of this type of risk. Due to the increase of natural disasters, urbanization, or climate change, the mutualization of the CatNat claims prevents uninsurability thanks to the CCR legal reinsurance. However, the performance of our model may lead to determining a segment for which the insurer has a higher negative margin. Even if, an important evolution of the French Cat-Nat program has been voted on the 16 December 2021, it might not be sufficient to overcome the insurability issue (Subsection 4.3) and even may jeopardize it.

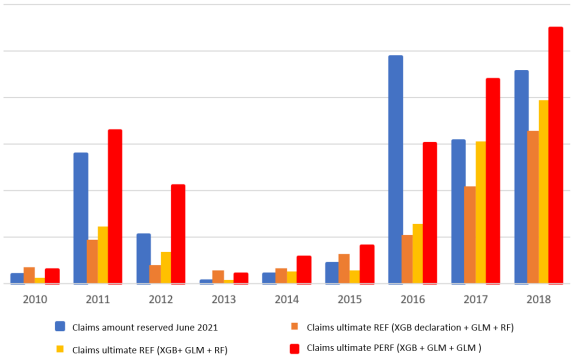
4.1. Reserving

Using meteorological variables, it is possible to estimate the global claim amount. In this paper, the claim reserve of the insurer from 2001 to 2020 was disposable. However, only the addresses' portfolio since 2015 was available. For the year 2010 to 2014, let's assume that the 2015 portfolio is a proper approximation for the 2010 to 2014 portfolio. This approximation is a quite good one in terms of exposition and portfolio stability. For non-geocoded addresses, the mean department values for the frequency and the claim amount is used. For the year 2012, it seems that the drought appeared in a zone well geolocated, leading to an overestimation due to the recalibration.

Results can be shown in Figures 14a and 14b. For recent years 2019 and 2020, the drought indexes are not available, and not enough municipalities have declared a subsidence CatNat. In recent years, there is an important difference due to the opening reserving process and the development of claims, the observed sinistrality is underestimating the global claims



(a) Prediction of the number of claim for reserving purposes. As expected, the year 2016 is abnormal. The R-squared metric equals 48% and 92% respectively for the referent model and the performant model.



(b) The comparison between the predicted ultimate claims amount and the cost (payment and reserving) calculated in 2021. The R-squared metric equals -21%, 43% and 69% for respectively the referent model using the XGB declaration + GLM ref + RF, the referent model using the XGB + GLM + RF and the performant model XGB + GLM + GLM.

amount significantly around 10 to 20% for recent years. Finally, claims are observed as if 2021 using the construction index from the FFB *fr. Fédération Française du Bâtiment* French Building Federation. The results are acceptable, but are difficult to validate objectively. The year 2016 is highly underestimated in frequency and in claims and the other way round, the year 2017 is overestimated in frequency. Comparison between the claims' amounts - Figure 14b is more difficult. Indeed, recent years' claims are still open and the amount for open reserving is underestimating the mean claim amount. Keeping in mind the data particularity, the R-squared metrics show that the reserving process seems better evaluated using the building geolocalisation information. As [5] concludes, the subsidence coverage is still difficult to model even when adding new information because difficult to validate properly.

4.2. Prevention and uninsurability

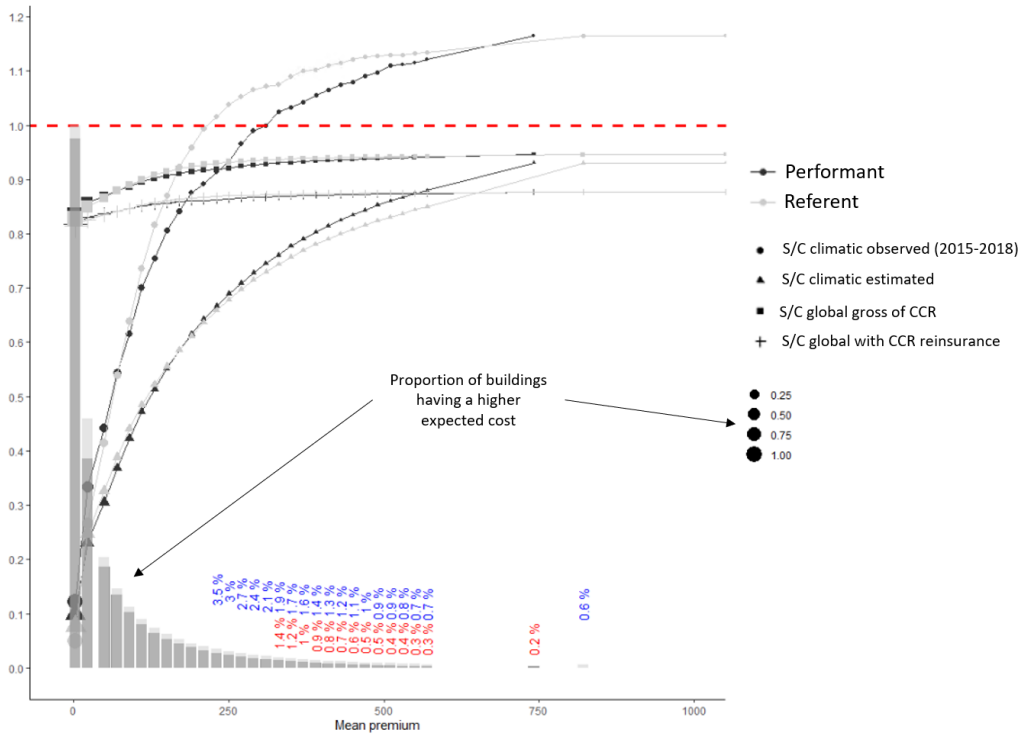
In France, the premium associated with CatNat coverage is enforced. Not knowing the global premium, let's assume the premium for damage coverage equals 300 euros, the CatNat premium is calculated as 12% of 300 euros[18]. In exchange for the CatNat tax/fee/contribution, the CCR reinsurer is providing a Quote Share Treaty up to 50% and an unlimited Stop-Loss starting around 300%-400% of the CatNat premium. Figure 15 compares the results between the performant model (GLM + GLM + XGBoost) and the referent model (RF + GLM + XGBoost) for pricing purposes, *i.e.* annual indicators are not used. Each line corresponds

[18]For houses, the mean premium (houses + flat) is around 255 euros in 2020 according to FA (the French Insurance Federation) information. The mean premium for individual houses is around 300-400 euros.

to S/C calculated on all contracts with a mean subsidence cost prediction or premium $P_i^{Subs.}$ lower or equal to the threshold u, *i.e.*

$$\frac{\sum_{i|P_i^{Subs.}<u} S_i}{\sum_{i|P_i^{Subs.}<u} C_i}, \tag{3}$$

where S_i is the claim cost of the policyholder and C_i its premium paid.



S/C	□	△	+	○
C	CatNat Premium	CatNat Premium	CatNat Premium + Damage premium	50 % CatNat Premium + Damage premium
S	Observed claims (2015-2018)	$E(\hat{S}) = P^{Subs.}$	$E(\hat{S})$ + 80 % Damage premium	50% $E(\hat{S})$ + 80 % Damage premium

Figure 15: Comparison between the referent model (RF + GLM) and performant model (RF + GLM) for pricing purposes. The histogram corresponds to the proportion of contracts for which the expected subsidence cost/premium is higher. The first performant bar is not equal to 1 because the residuals contracts correspond to houses having some missing information (period of construction or vegetated surface variable). Here, the S/C for a given premium represents the S/C of all contracts with a premium inferior or equal.

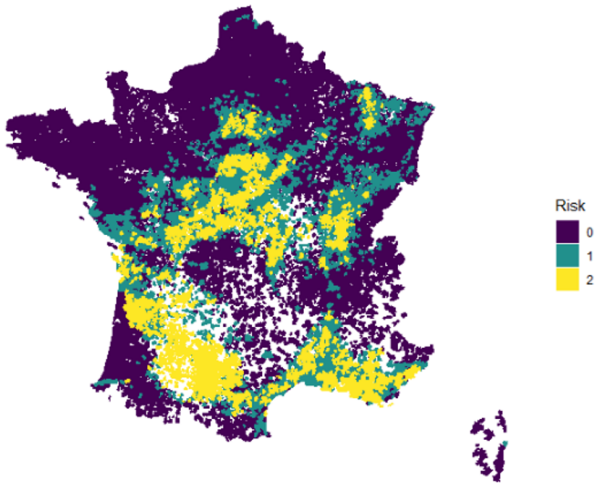
As explained in the previous section, the geolocated data is highly improving the model's performance, especially for the conditional frequency model. The differences between the referent model and the performant one show that insurers are now able to determine houses for which losses highly exceed the premium CatNat. One must also keep in mind that even if the French insurers must incorporate 12% of the sum of damage insurance coverage for CatNat protection, another way around is to not insure some houses (or do less competition or less discount for these houses, higher premium increase, ...). This could have an impact at all levels, *e.g.* on the claim process such as prevention during drought to very vulnerable houses with financial purposes. For instance, the CCR modulates the commission levels around 1% for insurers, which do proper prevention or indemnity controls to prevent fraud on climatic risks.

Studying the results, the subsidence impact is very important in the last few years. Remind that the 12% contribution concerns also floods, earthquakes ... which are, here, not considered in the S/C. In the last years, the cost of these climatic risks was low but historically the flood damage alone had a higher cost than subsidence. The CCR highly improves the S/C of reinsurers through his co-insurance. Yet, the increase of CatNat cost and variability increases the reinsurance cost. With the CCR and the mandatory CatNat program, 50% of the natural disaster losses are mitigated for all segments.

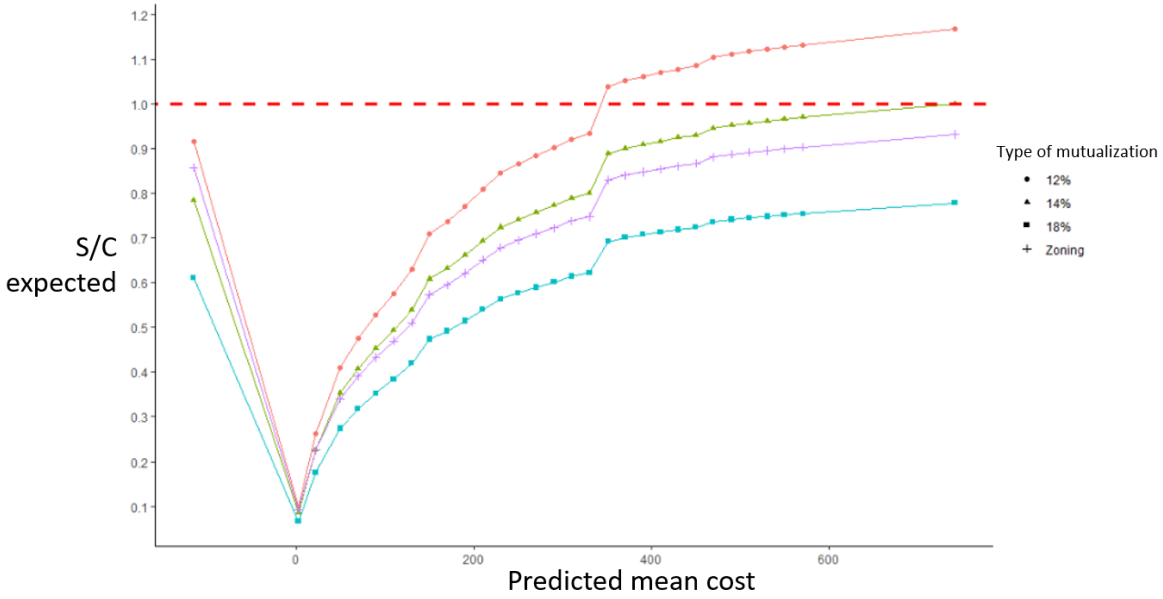
Are subsidence risks still insurable? From the insurers' point of view, even if the 4 consecutive years have a negative S/C climatic and have highly impacted the global S/C, insurers have still a global positive gross margin. Using the performant model, the insurers could refuse to insure 1.4% of the portfolio against 3.5% (referent model with the meteorological data) to have the same improvement of expected climatic S/C. Refusals are more doable, leading in our case to an improvement of around 20 bp of the expected climatic S/C without lowering too much the turnover. Created in 1958, the French administration *Bureau Central de Tarification*(BCT) has the task to help people who are not able to find an insurer for compulsory insurance. In such a case, this administration can impose a contract with a given premium to an insurer. For sure, the BCT can handle a small volume of demand, but a systemic rejection would be problematic and one would need to know the existence of such an administration.

4.3. Evolution of the CatNat program

On 28 December 2021, a CatNat reform has been enforced. The reform increased the delay during which a municipality can declare a CatNat and the delay during which an individual can declare subsidence damages from 2 to 5 years. This might increase the probability of CatNat and also the conditional frequency of claims compared to our model. Plus, French insurers must pay the costs of emergency relocation of disaster victims, as well as the costs of the architect and project manager. In the reform, the different deductibles depending on if the municipality has a prevention plan or not are unified.



(a) An example of subdividing the CatNat contribution according to drought expected cost on the municipalities available in our data. The CCR could produce an exhaustive map with the claims of the market.



(b) The combined ratio depending on the strategy with a contribution equal to 12%, 14%, 18% of damage premium or differentiated by zone. The first point refers to building for which the cost has not been predicted due to missing values.

Figure 16: The risk map according to drought exposure and vulnerability.

According to our results, laws enforced are heading in the right direction for the insurers' contributions to climatic risk management but might have a negative outcome. For sure, the CatNat tax/fee is not enough in front of the increase of drought frequency, climate change, or the lack of drought prevention on the structure during building construction. The gross cost and management cost will increase, especially on risky segments. To be still insurable, the expected S/C climatic should be lower than one minus the management cost ratio, around 10 to 20%. Increasing the cost will increase the S/C of the riskier segments only. In the future, if the subsidence cost does not decrease, insurers are now able for subsidence drought to focus on the most vulnerable houses to enforce some prevention plan. Nonetheless, it is more likely that some insurers may reluctantly insure houses vulnerable to subsidence risk. In one of the climatic ORSA scenarios of the ACPR, it is proposed to increase of the CatNat fee, up to 18 % by example. However, it seems better to adapt the CatNat premium depending on the expected cost while keeping the mutualization process. For instance, the lower risks could still contribute up to 12 %, the medium-risk up to 14% and the higher up to 18% as shown in the risk maps - Figure 16a - and would equilibrate - Figure 16b - the combined ratio. The idea is to keep the mutualization process without penalizing areas without CatNat risks. Moreover, a municipality could change its' zone risk if sufficient anticipation and risk's annihilation are set up.

This would increase the premium but lower the financial gain of insurance refusal. Finally, this process should also be done at least for floods. Some discussions on *Taxonomie1* of EIOPA (European Directives) are taking about including prevention measures in the design of products about rewards and duty of advice and asking for indicators for pricing and modelling of climate risks disasters. The brand image of these indicators is a powerful tool in our point of view if they are refined enough.

In short, the French CatNat program has proven its robustness since its creation in 1989 for subsidence risk. As researchers, the CatNat reform is heading in the right direction, better protecting the insured person. Nonetheless, the increase of the CatNat premium conceals the true problem. We recommend focusing more on risk prevention and annihilation by promoting actions against subsidence risks, homogenous risks' portfolios through taxes or lower reinsurance contracts so that these risks would be still financially attractive. Should also the building industry not only insurers participate in the CatNat program ?

5. Conclusion

This paper shows that the improvement and transparency of open data available in France allow actuaries to model subsidence claims more precisely. To access this information, insurers should geolocate each building insured. The downside of the model improvement is that some houses may become uninsurable. Even if a better prevention plan could be done, the

authors are afraid that the insurance market prefers to not accept a non-negligible amount of vulnerable houses. French and European legislators are heading in the right direction by increasing the insurers' CatNat contribution and prevention attractiveness. Nonetheless, the CatNat premium should be increased and further regulation should be done at the root of the risk (during the construction or during drought) in order to increase the insurability of this type of risk for all segments.

This work shows the discrepancies between Machine Learning methods (XGBoost and Random Forest) and GLM. The linear structure of GLM versus ML non-linear one embodies why French actuaries prefer GLM for most actuarial applications. Indeed, the claims' data are neither perfect nor fully developed. Therefore, a black box method often leads without any proper control to a fully data-driven model without fully understanding the underlying particularity of data designs. This paper shows an application dealing with the robustness of black-box methods. Still, a lot more controls/tests need to be developed for actuarial uses.

Acknowledgement

This paper stems from a mission of research of development in the context of the product "Smart Home Pricing". The authors thank the anonymous firm which has allowed us to use the portfolio, geolocate it and create this data set. All the values given in this paper have been anonymized. The views expressed in the paper do not represent the views or positions of the different firms linked to this project.

References

[1] Christoph Bergmeir, Rob J Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 2018.

[2] Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. mlr: Machine learning in r. *Journal of Machine Learning Research*, 17(170):1–5, 2016. URL `https://jmlr.org/papers/v17/15-066.html`.

[3] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] Eugénie Cazaux, Catherine Meur-Férec, and Cédric Peinturier. Le régime d'assurance des catastrophes naturelles à l'épreuve des risques côtiers. aléas versus aménités, le cas particulier des territoires littoraux. *Cyber geo, european journal of geography*, 2019. doi: https://doi.org/10.4000/cybergeo.32249.

[5] A. Charpentier, M. R. James, and H. Ali. Predicting drought and subsidence risks in france. *Natural Hazards and Earth System Sciences Discussions*, 2021:1–27, 2021. doi: 10.5194/nhess-2021-214. URL https://nhess.copernicus.org/preprints/nhess-2021-214/.

[6] Johnny Douvinet and Freddy Vinet. La carte des arrêtés catnat pour les inondations: analyse spatio-temporelle. *M@ppmonde*, 1, 2015. URL http://mappemonde-archive.mgm.fr/num35/articles/art12301.html.

[7] Michael Alin Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203, 1960.

[8] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[9] Andrés Hoyos-Idrobo, Yannick Schwartz, Gaël Varoquaux, and Bertrand Thirion. Improving sparse recovery on structured images with bagged clustering. In *2015 International Workshop on Pattern Recognition in NeuroImaging*, pages 73–76. IEEE, 2015.

[10] M. Ionita and V. Nagavciuc. Changes in drought features at the european level over the last 120 years. *Natural Hazards and Earth System Sciences*, 21(5):1685–1701, 2021. doi: 10.5194/nhess-21-1685-2021. URL https://nhess.copernicus.org/articles/21/1685/2021/.

[11] Spinoni Jonathan, Naumann Gustavo, Vogt Jürgen, and Paulo Barbosa. Meteorological droughts in europe: Events and impacts – past trends and future projections. Technical Report EUR 27748 EN, Publications Office of the European Union, Luxembourg, 2016.

[12] Thomas Mack. A simple parametric model for rating automobile insurance or estimating ibnr claims reserves. *ASTIN Bulletin: The Journal of the IAA*, 21(1):93–109, 1991.

[13] Guillaume Maillard, Sylvain Arlot, and Matthieu Lerasle. Aggregated hold-out. *Journal of Machine Learning Research*, 22(20):1–55, 2021. URL http://jmlr.org/papers/v22/19-624.html.

[14] Thomas B McKee, Nolan J Doesken, John Kleist, et al. The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, volume 17, pages 179–183. California, 1993.

[15] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

[16] Arthur E Renshaw and Richard J Verrall. A stochastic model underlying the chain-ladder technique. *British Actuarial Journal*, 4(4):903–923, 1998.

[17] Jonathan Spinoni, Gustavo Naumann, Hugo Carrao, Paulo Barbosa, and Jürgen Vogt. World drought frequency, duration, and severity for 1951–2010. *International Journal of Climatology*, 34(8):2792–2804, 2014.

[18] Jean-Philippe Vidal and Steven Wade. A multimodel assessment of future climatological droughts in the united kingdom. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 29(14):2056–2071, 2009.

[19] Donald A Wilhite and Michael H Glantz. Understanding: the drought phenomenon: the role of definitions. *Water international*, 10(3):111–120, 1985.

[20] Guangyong Zou. A modified poisson regression approach to prospective studies with binary data. *American journal of epidemiology*, 159(7):702–706, 2004.

Conflicts of interest

The first author declares a conflict of interest, being a full-time salary of the project commercially named *Smart Home Pricing* developed by Addactis FRANCE and Nam.R. Nonetheless, he declares no financial interest. The remaining author has no conflicts of interest to declare.

Appendix A. Gaspar Database

The quality of the GASPAR database is good starting from 2005. Based on current knowledge, no proper quality evaluation has been done on drought and subsidence CatNat declaration and PPRN. Through extensive analysis, few errors exist and appear only before 2009. The database version studied is from 19 July 2021. The subsidence PPRN is associated with the number 157 but some PPRNs -134 (drought) also refer to it. The following municipalities PPRN have been corrected ”09317”, ”24013”, ”31438”, ”31307”, ”32168”, ”36151”, ”36183”, ”36201”,”68164”, ”68310”, ”81091”, ”81113”, ”81153”, ”81155”, ”89457”,”89275”, ”89213”, ”32022”, ”32081”, ”33063” and ”89356”. As shown on Figure A.17, the CatNat declaration was not as clear before 2001. Therefore, the quality of all the CatNat declarations before 2001 should ideally not be taken into account.

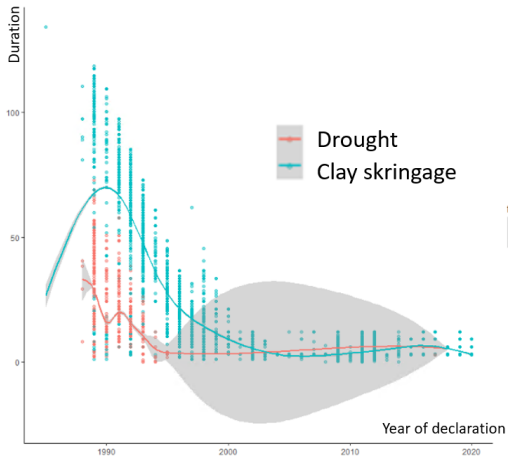


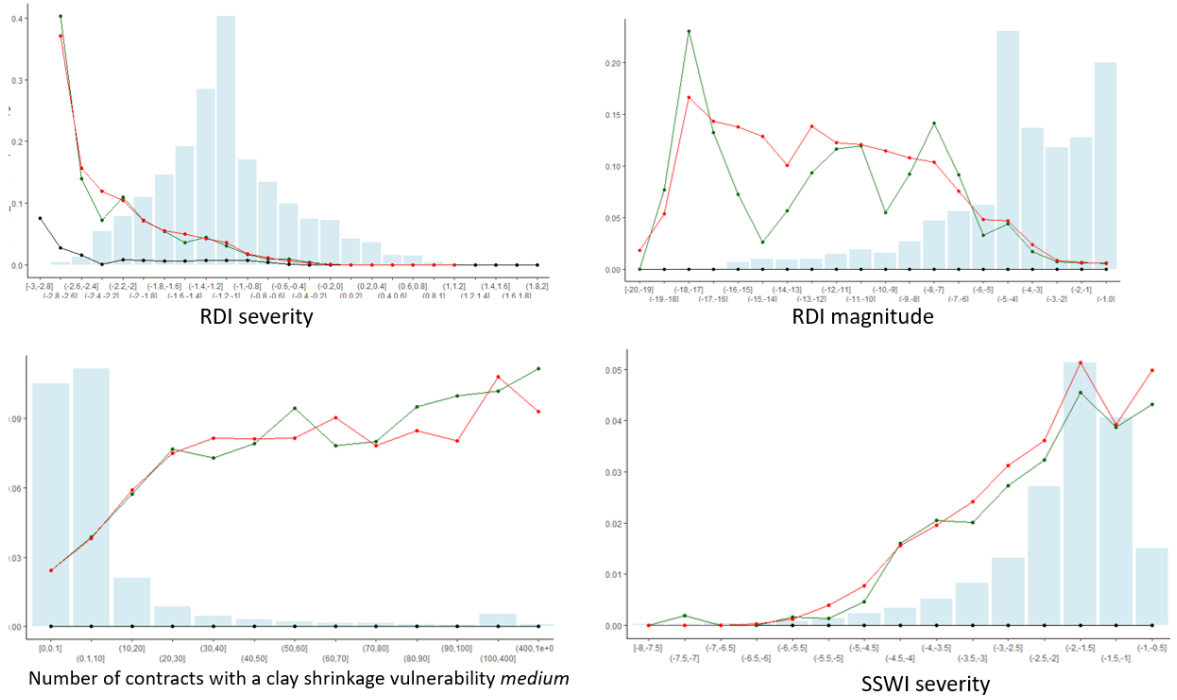
Figure A.17: The data used comes from the gross GASPAR database (19 July 2021). The subsidence/drought CatNat process starts in 1989 and the duration stationarity of the data starts at the end of 2003. At the start, the criteria for clay shrinkage (subsidence) and drought were not clear enough. To see the trend, for each type of CatNat, a line corresponding to the marginal impact of an univariate GAM regression is plotted.

Appendix B. One-way analysis

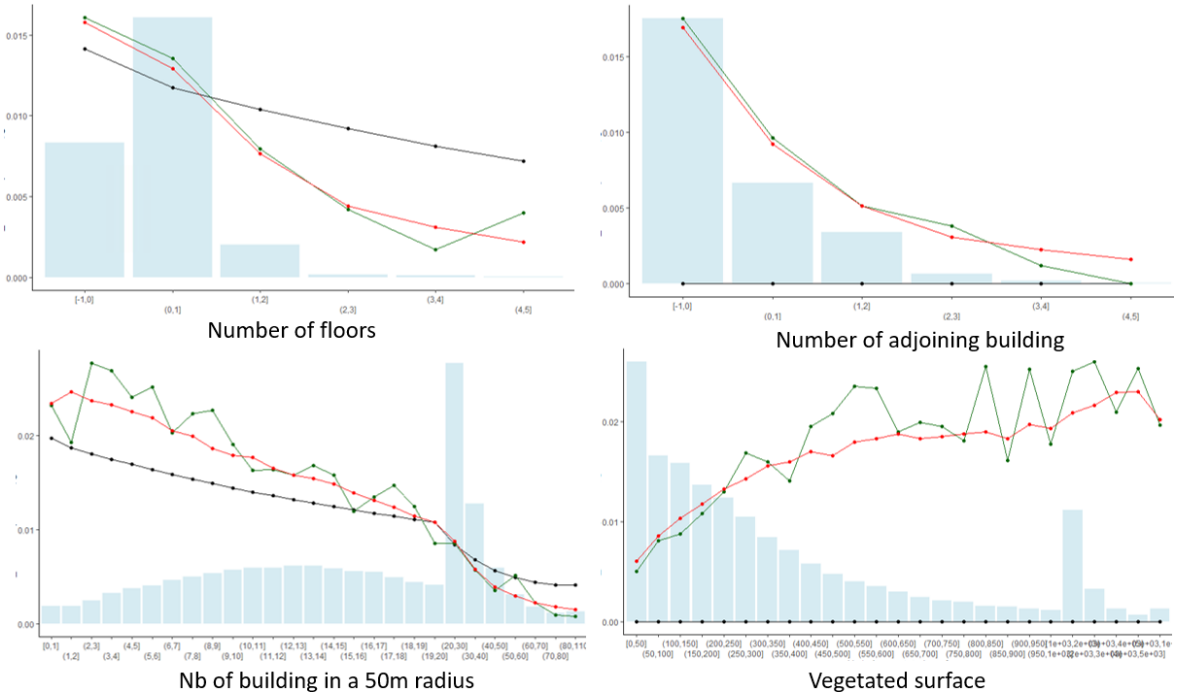
The following Figures B.18a and B.18b are some one-way analysis of variables used into GLM evaluated on the testing database. Blue histograms refer to the exposition, the green line to the observed value, and the red one to the predicted value. The black line refers to the marginal impact captured by the GLM. For purposes of confidentiality, not all marginal impacts are plotted for confidentiality purposes.

Appendix C. Some Shapley interactions

The following Figures C.19 and C.20 are Shapley values calculated on a simple XGBoost model trained on 2010-2018 historical data including the meteorological annual indicators. The subset used to determine the Shapley values is 20 000 random rows from the training and the testing dataset. The Shapley score convergence is stable starting from 15 000 rows used.



(a) One way-analysis for GLM CatNat.



(b) One way-analysis for GLM conditional frequency.

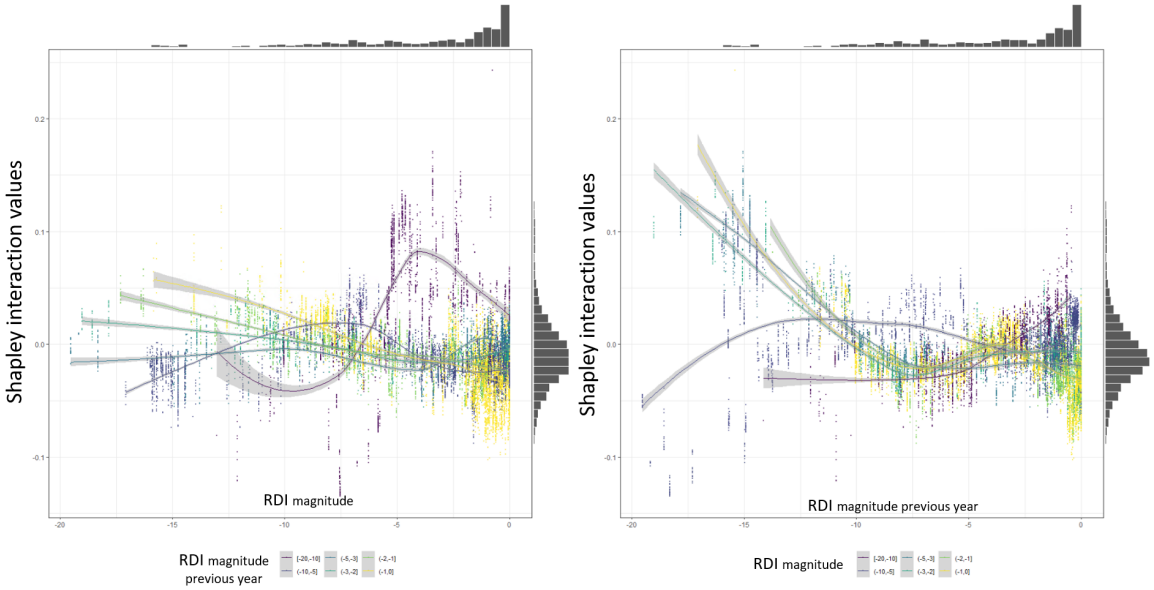


Figure C.19: Shapley interaction plot shows links between the previous meteorological year and the following. Each line represents the marginal impact of a GAM model fitted by subcategories of the other variable.

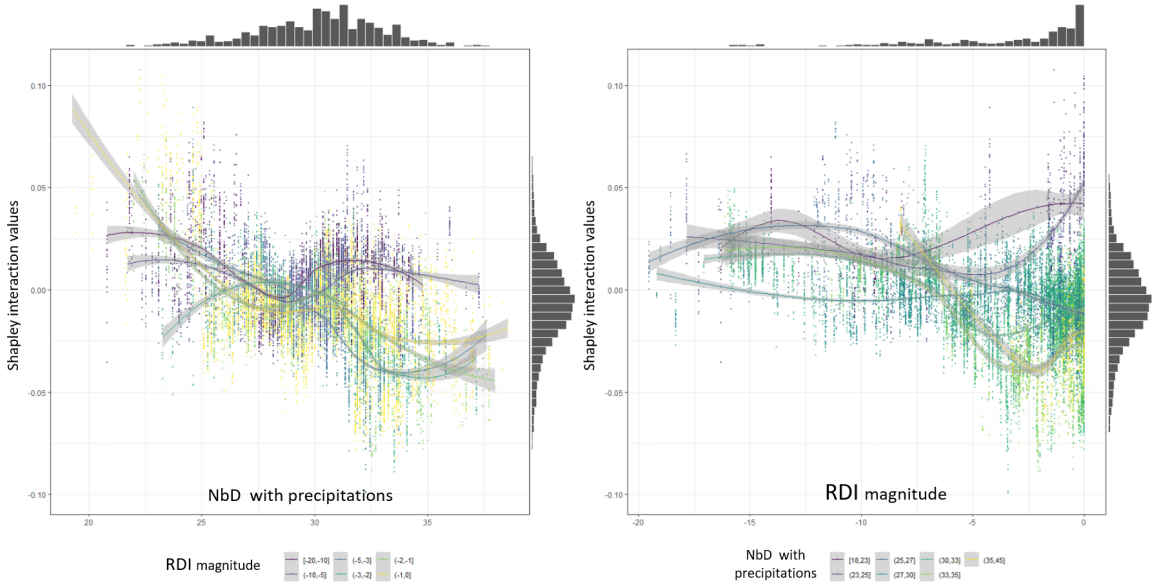


Figure C.20: Not all interactions are well captured. The variable precipitation quantity significantly improves the model's performance, but the marginal impact is different depending on the municipality's climate. Other interactions with the indicator SSWI or with the precipitation quantity of the previous year are also relevant.