



HAL
open science

Application of machine learning methods for cost prediction of drought in France

Antoine Heranval, Olivier Lopez, Maud Thomas

► **To cite this version:**

Antoine Heranval, Olivier Lopez, Maud Thomas. Application of machine learning methods for cost prediction of drought in France. 2022. hal-03310875v2

HAL Id: hal-03310875

<https://hal.science/hal-03310875v2>

Preprint submitted on 2 Feb 2022 (v2), last revised 2 Aug 2022 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Application of machine learning methods for cost prediction of drought in France

Antoine HERANVAL^{1,2*} · Olivier LOPEZ² ·
Maud THOMAS²

Received: date / Accepted: date

Abstract This paper deals with the prediction of the total amount of the damages caused by a drought episode under the French “Catastrophe Naturelle” regime. Due to the specificity of this regime, an early prediction of the final amount of an incident is particularly strategic. Thanks to a partnership with the Mission Risques Naturels, we had access to a database of natural catastrophe claims fed by the major French insurance companies. Combining the information about drought event claims contained in the database with meteorological and socioeconomic data allowed us to increase our knowledge of the exposure. Our prediction approach relies on the comparison of different statistical models and machine learning algorithms. To improve the performance, we propose an aggregation the different models. The main difficulty comes from the fact that the data is imbalanced since a large majority of cities were not stroke by a drought event. The predictions were thus assessed by F1-scores and Precision and Recall curves.

Keywords Natural Catastrophe · Generalized Linear Models · Lasso and Elastic-Net penalties · Extreme Gradient Boosting · Random Forests

Mathematics Subject Classification (2020) MSC code1 · MSC code2 · more

1 Introduction

According to the Fédération Française des Assurances, FFA (French Federation of Insurance Companies), the cost of the damages caused by natural catastrophes, such as drought, is expected to increase in the following years in France [3]. Although not making the front page, the risk of drought is responsible for about 30% of the total amount of claims paid by the French regime CatNat (Régime d’indemnisation des catastrophes naturelles) [5]. This represents over 11 billions euros with some extreme years, like 2003 where the cost rises up to 2 billion euros. The rise of the number of drought events is apparent: out of the six most costly events since the creation of the regime in 1982, three have occurred after 2010 [5]. This is mainly due to the general growth of wealth in France and to climate change. Climate change has an important

¹Mission Risques Naturels, 1 rue Jules Lefebvre 75009 Paris, France

²Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, 75005 Paris, France

*Corresponding author

E-mails : antoine.heranval@sorbonne-universite.fr, olivier.lopez@sorbonne-universite.fr, maud.thomas@sorbonne-universite.fr

impact on drought and its effects, especially on individual houses due to clay-related subsidence. Clay-related subsidence is caused by clay shrinking and swelling in response to wetting and drying conditions. Indeed, because of the volumetric changes in soil mass, clay shrinking and swelling cause vertical and horizontal ground movements, which can lead to significant damages to infrastructure and especially to individual houses [2, 12]. These damages have been reported in other countries as well and the associated cost is also very large, up to £500 millions per annum in the UK, for instance [26].

In this paper, we propose a methodology to estimate the total amount of the consequences of a drought event shortly after its occurrence for the entire French market. The main goal is to provide tools for insurance companies to assess the severity of future drought events. Because of the large incurred amounts, the potential involvement of the government, or the quality of prevention and assistance delivered to policyholders, evaluating the order of magnitude of the cost of such an episode is of most importance and a challenging task.

The specificity of the French system, under the “Régime Catastrophe Naturelle” (CatNat regime), makes it even more difficult. Indeed, a very specific administrative procedure—involving insurance companies and the government—governs the compensation and the management of such claims. Financial assistance is then triggered by a decree, published only after an examination process that can take a significant amount of time. The approach we develop in this paper aims at predicting whether a geographic area (namely a city) will be affected by a drought event. The knowledge of such information is of course crucial to determine the amount of reserves required to face such an event. Once a prediction has been made, a cost can be estimated from the number of exposed houses.

Our methodology to predict the cost of drought events relies on the comparison of different statistical methods such as Generalized Linear Models (combined with Lasso and Elastic-Net penalties [21]) with machine learning algorithms, such as Random Forests [13, 33] or Extreme Gradient Boosting [17]. The calibration of these methods is performed on a large database provided by Mission Risques Naturels, covering approximately 70% of the French non-life insurance market. An important difficulty stands in the fact that this database is very imbalanced. Indeed, catastrophic events such as drought are relatively rare, and most geographical areas are not affected by any claim at all. To improve the performance and to benefit from all the models considered, we propose an aggregation of the outcomes on which we can base new predictions. The predictions obtained from the different models are thus assessed with Precision and Recall curves, F_1 –scores and confusion matrices.

Among recent works related to the evaluation of the impact of drought in France, Ecoto et al. (2021) [20] uses a Super Learner methodology to predict the cost of a claim at city level. The prediction problem is not identical to ours, since the state of natural catastrophe of the cities that are studied in [20] is already known at the time of the prediction. In our case, this information is unavailable and has to be predicted: since we are interested at the evaluation of the cost at national level, the estimation of the probability that a city will benefit from the natural catastrophe status is an important issue. Charpentier et al. (2021) [16] consider a similar problem to ours, but relying on different data and indices to measure the severity of the drought. All of these approaches, including the one we propose in the present paper, contribute to an evaluation of the impact of drought which today lacks of precise official indicators, except for the one produced by Caisse Centrale de Réassurance, and Mission Risques Naturels (which mission consists in aggregating the experience of the main actors of this sector and to develop a collective expertise of this risk).

The rest of the paper is organized as follows. In Section 2, we describe the framework of this problem and the variable used to predict the cost. Section 3 is devoted to the general description of the statistical models used for the prediction. In Section 4, these models are applied to a real

database. Section 5 presents the results of the cost prediction. The paper ends with a discussion in Section 6.

2 Description of the problem and associated data

This section presents the context of this work and the problem related to the French CatNat regime. A short presentation of this regime is made in Section 2.1. We then describe our classification problem in Section 2.2, which consists in predicting that a city will be affected by a claim. The database, developed by Missions Risques Naturels (MRN), used to perform this prediction is described in Section 2.3. The covariates are presented in Section 2.4. We particularly focus, in this last section, on a spatial-temporal meteorological index, the Standardized Soil Wetness Index (SSWI), used to characterize the intensity of the drought. The propensity of the soil to clay shrinking and swelling relies on the cartography produced by the Bureau de recherches géologiques et minières (BRGM), a French geological and mining research institute [32]. Finally, we summarize the whole methodology in Section 2.5.

2.1 Short description of the specificity of French CatNat regime

In France, natural catastrophes are insured through a public-private partnership, called the CatNat regime. This French specificity strongly influences natural catastrophe claim management. This natural disaster compensation scheme was created by the Law of July 13 1982, and is based on a solidarity principle: for every contract, the same additional premium insurance rate, fixed by the government, is used to compensate for the losses of natural disasters. The scope of its application is large: for example, it covers floods, mudslides, earthquakes and landslides. However, storms, hail and snow are not included. Without going into the functional details of this regime of compensation, it is important to note that before receiving the compensation, a government decree, recognizing a city in the state of natural catastrophe or not, must be published in the “Journal Officiel”, where all laws and legislative events of the French Republic are published. To receive a compensation, the policyholder must make a request to the CatNat Commission.

This decree comes after a request from the mayor of the city asking to the government to recognize the event as a natural catastrophe. The decree is motivated by an inter-ministerial commission, which assesses the exceptional situation of the event at city level. For drought events, the evaluation is based on the soil type and moisture. This corresponds to the exposition of clay shrinking and swelling (soil type) and the meteorological intensity of the drought in the city (moisture). The classification of the propensity to clay shrinking and swelling is publicly available through a fixed cartography produced by the BRGM shown in Figure 2 [32]. To measure the wetness of the soil, the Commission uses an indicator developed with the French meteorological institute Météo France. This indicator of the clay risk factor consists of several geological indices obtained from experts. Based on the values of this index for several months and on whether the city has clay areas, the decree will recognize the city in a natural catastrophe state [7,8,9]. This process can be time-consuming, the average time between the occurrence of the drought event and the decision is about 18 months [5], which is a long time for both the policyholder and the insurer. Moreover, if no request is made, or if the Commission refuses the request, there will be no compensation from this scheme. In this case, additional coverage can be provided by the insurance company but for the clay-related subsidence in France, it is very rare. Considering the large amounts at stake and the long delay, insurance companies are trying to anticipate the total cost of such events. Estimating the cost of a drought episode soon after its occurrence is thus of most importance.

2.2 A binary classification problem

A first step to predict the cost of a drought event would be to identify the cities that might be recognized in a natural disaster state. Unfortunately, this is a difficult task since firstly, there is an uncertainty on whether the city will make a request to the Commission and secondly, the meteorological index used by the Commission as one of the criteria is not publicly available early enough after the occurrence.

To overcome this issue, we propose to rather predict the cities that may have a claim, since thanks to a partnership with the MRN, we have access to a database containing the past claims that have occurred in France.

Mathematically speaking, we are dealing with a binary classification problem. Let $Y \in \{0, 1\}$ denote the response variable and $X \in \mathbb{R}^p$ the covariates, $Y = Y_{ij}$ is equal to 1 if a drought event has occurred in city i in year j and 0 otherwise. Our goal is thus to estimate $\mathbb{P}[Y = 1 | X]$. The results of this prediction problem are then linked with a cost in Section 2.5. In the next sections, we describe the database and the covariates used to address this problem.

2.3 The SILECC database

Thanks to a partnership with the MRN, we have access to the database of the claims related to climatic and natural catastrophe in France (SILECC) [4]. This database covers about 70% of the French non-life insurance market by aggregating the claims of 12 major French insurance companies. This database records the natural disaster claims in France from 1987 to 2018. Every claim has been standardized and geolocalized. It is very useful for the insurance market and the FFA since it enables the tracking of the claims related to natural hazards and where they have occurred. It is also of general interest, e.g. to update the cartography of clay shrinking and swelling, as mention in Section 2.4.

While the database covers several natural hazards, we focus in this paper on drought events, that is claims related to clay shrinking and swelling. Since some companies have not contributed for the whole period, we considered to focus on the period from 2003 to 2018 for which we had a sufficient number of claims. This period provides strong representativeness of drought events in France and covers major episodes such as the ones observed in 2003 and in 2011. We used the data from 2003 to 2017 for estimation and 2018 was kept as an illustration of our prediction methods.

In the database, the number of cities affected by a claim represents 6% of the total number of cities in mainland France. This corresponds in average to 1 948 cities with a claim per year, out of 34 840 cities in mainland France. This ratio varies during our time period. As written before, for some years, a large number of claims was observed, such as in 2003 where 25 % of the total number of cities were affected by a claim, 2017 and 2011 with 12 % and 10 % are also important. Figure 1 shows the yearly percentage of cities affected by a claim out of the total number of cities affected by a claim between 2003 and 2017.

2.4 Covariates

To characterize the propensity to clay shrinking and swelling in the soil, we rely on an indicator based on geological expertise which is used in the cartography published by the BRGM [32]. This index is a complex aggregation of characteristics related to the lithology (measuring the formation and the proportion of clay), to the mineralogical composition of the clay area (with a special focus on the proportion of smectites which are of particular importance), and to results

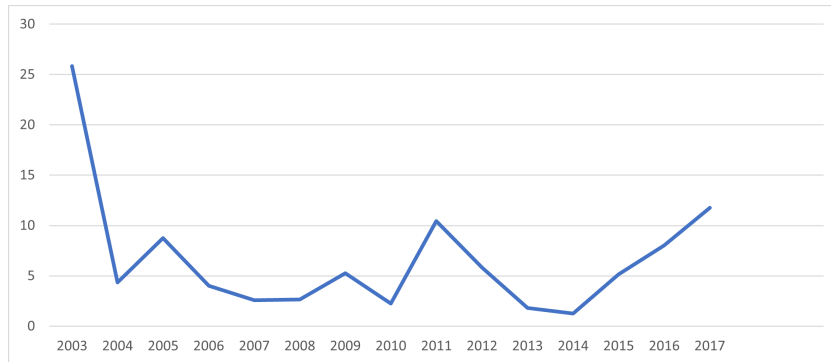


Fig. 1 Yearly percentage of cities affected by a claim out of the total number of cities affected by a claim between 2003 and 2017

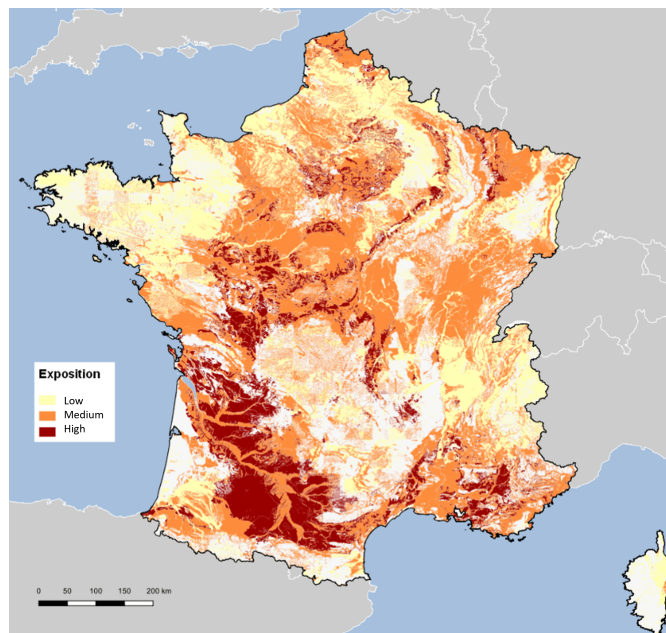


Fig. 2 Cartography of the propensity of clay shrinking and swelling clay in France. Source: BRGM [32]

of tests performed by laboratories on claims that have occurred in a given region. Historical claims are also used to reflect the frequency of incidents in the area. Updating this index has been done with the help of MRN in 2019, as described in [6]. Once computed, this index can be understood as a risk factor, and a ranking of the different areas is performed, defining three hierarchical classes, providing a national cartography (at city level) that describes the propensity to clay swelling. This cartography is shown in Figure 2 from [32].

This cartography allows us to compute the surface and the proportion of each zone (low, medium and high) at city level. We then estimated the number of individual houses in each zone using the data of INSEE in 2015 (French public statistical organization). To take the evolution of the number of individual houses into account, we applied an augmentation or reduction of 1% for each year [11].

Regarding the meteorological index, the one used by the Commission has been made available since April 2021 but will not be published early enough for cost prediction. In this respect, we used another spatial-temporal meteorological index, produced by Météo-France, the Standardized Soil Wetness Index (SSWI), as an indicator of the severity of a drought event. This index comes from a research project of Météo-France called Climsec, described in [31]. The calculation of the SSWI is done through the analysis of precipitation, soil moisture and streamflow outputs from the Safran-Isba-Modcou (SIM) hydrometeorological suites [22], and inspired by the Standardized Precipitation Index (SPI) computation procedures [24]. The description of the SSWI is beyond the scope of this paper, a description can be found in [30]. Four time series are then obtained from the SSWI time series as moving averages over one, three, six and twelve months. This gives us four indices for each month representing the wetness of the soil.

The SSWI is a standardized index, takes thus values centered around 0. A negative value suggests drought whereas a positive one suggests wetness. Figure 3 shows the geographical distribution of the SSWI for 2018. We can see that it is highly variable and that 2018 was a year with an important drought in France.

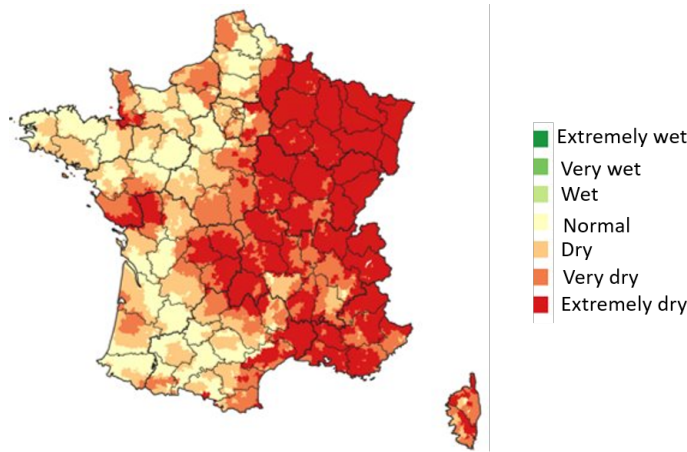


Fig. 3 Cartography of the SSWI for the year 2018. ©MRN. Sources: Météo France SSWI, ADMIN EXPRESS (IGN)

We also computed four other indices to characterize the drought event itself, as defined in [30].

- the duration: the number of consecutive months during which the SSWI is negative;
- the severity: the absolute value of the minimal value of the SSWI reached during the event;
- the magnitude: the absolute value of the sum of the SSWI during the event,
- the rarity: a classification of the severity in 7 classes (Extremely wet, Very wet, Wet, Normal, Dry, Very dry and Extremely dry) as shown in Figure 4.

These indices are calculated for each city and for each year. In the case of multiple events (in our case the maximum is four events) in one year, we use the value of the indices for all events occurring during the year. Figure 4 illustrates the definition of these indices on one example.

We also used indications on the decree of natural catastrophe. Criteria leading to the decree of natural disasters have changed six times over the past 20 years. Therefore the same effects might not lead to the same consequences in our database, depending on the period. All of these

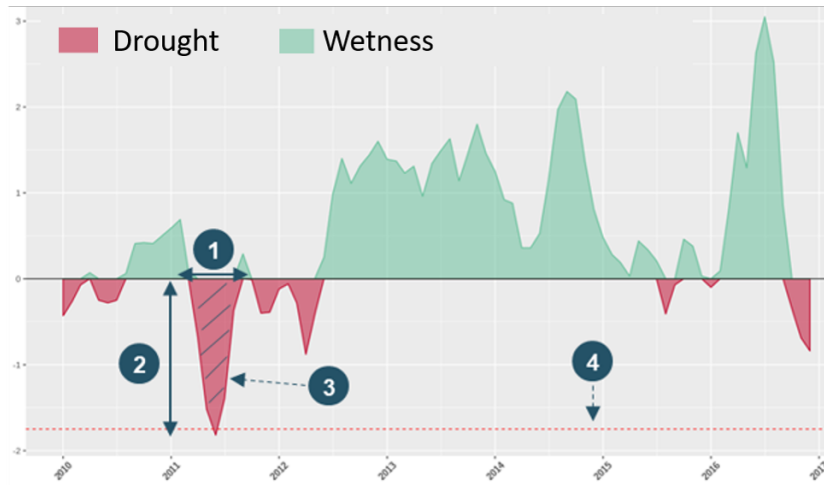


Fig. 4 Description of the four indices based on the SSWI we use based on one example. 1 represents the duration of the event, 2 its severity, 3 its magnitude and 4 its rarity.

constraints make the prediction delicate. We thus implemented a variable that indicates the criteria used by the Commission at the time of the claim. We also considered the results of the decision of the Commission as if they have used our SSWI meteorological index instead of theirs.

Overall our database contains 155 variables and 522 600 observations (all variables are numerical and categorical variables have been encoded as binary variables). A table with a description of these variables can be found along with the code on a Git. To summarize we have 96 covariates relative to the SWI (the minimum, and maximum of each index for each month), 37 variables describing the drought events, 4 for the criterion used by the Commission, 4 about the past declarations of natural catastrophe, 1 on the city population, 1 on its urban area, and lastly 11 for the propensity of shrinking and swelling of clay and for the number of houses in each area.

Those variables along with the database described in Section 2.3 constitute our learning database, on which machine learning models, described in Section 3, will be trained.

2.5 Overall methodology

The first step of our method is to predict the cities that will be affected by a claim during a drought event. For that, we used different machine learning models described in Section 3. Once we know the cities that are likely to be affected by a drought event, we calculated the number of houses in these cities that have a propensity to clay shrinking and swelling. To do that we used the cartography done by the BRGM (Figure 2), we counted the number of houses in the city with a propensity to clay shrinking and swelling, that is houses localized in a zone with medium and strong propensity. We then used linear regression to link the number of houses to the cost of the event. This linear model has been trained on our claim database. Figure 5 summarizes this overall methodology.

Let us write the cost for a city as random variable $Z = Y \times M$, where $Y = 1$ if the city has a claim (and 0 otherwise), and M is the amount of the claim, the classical approach typically assumes that Y and M are independent. This assumption may be questionable. A possibility to avoid it would be, for example to consider a more elaborate model for M as the one we developed below, for example linking M through a large number of covariates (so that Y and M may be

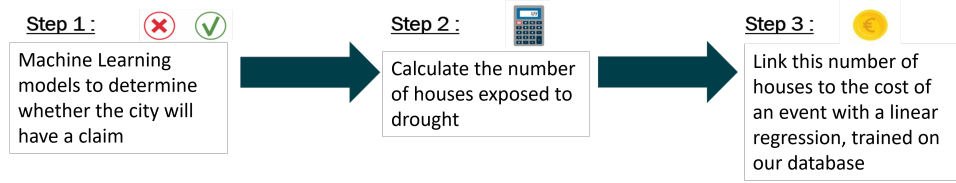


Fig. 5 Overall methodology.

independent but conditionally on the values of these covariates). However, this would require more information that we do not have at our disposal.

3 Statistical models

In this section, we describe the different models that we will use and combine to perform the prediction. The tuning are reported along with the code on Git. Section 3.1 is devoted to the Generalized Linear Model with elastic-net estimator, which is a parametric model adapted to covariates with high dimension. Random forests are described in Section 3.2, while a short presentation of boosting methods like XGBoost is done in Section 3.3.

3.1 Generalized linear model with penalty

The Generalized Linear Model (GLM), see for example [25] or [19], is a generic way to consider regression problems which is widely used in insurance. This class of models states that, for a response variable Y and $X \in \mathbb{R}^p$ some covariates,

$$g(E[Y|X]) = X\beta,$$

with $\beta \in \mathbb{R}^p$ is the vector of unknown parameters, and g some monotonous known function, called the link function. Additionally, the conditional distribution of Y given X is assumed to belong to some fixed family of distribution from the exponential family.

In a binary classification problem as ours, the distribution of $Y | X$ is assumed to be a Bernoulli distribution with unknown parameter $p(X) = \mathbb{E}[Y|X]$. Regarding the link function g , a standard choice consists in taking $g(y) = \text{logit}(y) = \log(y/(1-y))$. This corresponds to the canonical link function, the link function leading to the best theoretical properties in GLM. It is also a simple function that maps $[0, 1]$ into \mathbb{R} .

Estimation can be performed by maximizing the likelihood of the model. However, in our case, the dimension p of the covariates is relatively high. This creates a problem since the statistical precision diminishes with the number of coefficients to estimate. Moreover, many numerical issues can occur. On the other hand, most variables are likely to be irrelevant (but of course, one does not know which by advance). Hence, the GLM elastic-net estimator (GLMNET) is a way to reduce dimension by solving the numerical instability [34].

Let $f_\beta(y, x)$ denote the likelihood of the model. The GLMNET estimator is defined as

$$\hat{\beta} = \arg \max_{\beta} \sum_{i=1}^n \log(f_\beta(Y_i, X_i)) - \lambda \{ \alpha \|\beta\|_1 + (1 - \alpha) \|\beta\|_2 \},$$

with $\|\beta\|_1$ (resp. $\|\beta\|_2$) denotes the l^1 - (resp. l^2 -) norm of the vector β , the hyper-parameters λ and α being positive constant. The penalization of the log-likelihood by $\|\beta\|_2$ corresponds to a

Ridge penalization (see [23]), which stabilizes the estimation result by reducing some numerical issues that may happen in high dimension. On the other hand, the penalization by $\|\beta\|_1$ corresponds to a Lasso penalty (see [29]), designed to produce a sparse model, i.e. a model in which most coefficients of $\hat{\beta}$ are equal to zero. Hence it allows one to reduce the effective dimension of the covariates. The constants λ and α are chosen by cross-validation.

The advantage of GLMNET is to produce an intelligible and easily interpretable model. On the other hand, being able to automatically select the variables that have an effect on Y allows us to consider a model complex enough to expect a good fit. Nevertheless, the underlying parametric assumption may be too strong in practice. This is why we also turn towards “black-boxes” techniques from machine learning.

3.2 Random Forests

Random Forests (RF) constitute a machine learning procedure based on the aggregation of regression trees [13]. Regression trees, as introduced by Breiman [14], estimate the function $p(x)$ by

$$\hat{p}(x) = \sum_{j=1}^K p_j R_j(x), \quad (1)$$

where, for all x , $R_j(x) = 0$ for all j except for one. Namely, R_j are “rules” that are associated with a partition of the covariate space, allowing to affect x to the unique set of the partition to which it belongs. In regression trees, these partitions are hyper-rectangles, that is $R_j(x) = 1$ if and only if $x \in \cap_{k=1}^d \{x : x_{k,l} \leq x_l \leq x_{k,r}\}$, where $x = (x_1, \dots, x_d)$. This partition is obtained iteratively through the CART algorithm, see [14]. The estimation of the values p_j is then done for each region of the space $R_j(x)$.

Regression trees have many appealing properties, like allowing to introduce non-linearities while still producing a model which can be easily understood. However, their main drawback is their instability: new incoming data may disrupt the structure of the partition. RF are a way to stabilize this technique, while capturing more elaborate shapes of regression function $p(x)$. They are obtained by averaging regression trees with some specificities:

- each tree is of small size (a small value of K in Equation (1) is imposed);
- each tree grows on a separate bootstrap sample;
- the rules of a given tree are based only on a small subset of the covariates (subset which is selected randomly).

In the following, we use the implementation of RF from the R-package `ranger` [33].

3.3 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBOOST), see [17], is an alternative method to RF which also relies on regression trees, but instead of fitting these trees simultaneously, they are fitted iteratively. The predictor $\hat{p}^{(t)}(x)$ at the t -th step of the algorithm is obtained from the predictor $\hat{p}^{(t-1)}(x)$ at the $(t-1)$ -th step by $\hat{p}^{(t-1)}(x) + \pi_t(x)$, where $\pi_t(x)$ is a regression tree selected in order to make the loss function decrease as much as possible, that is to maximize the log-likelihood in the Bernoulli case with a regularization penalty.

Let $\ell(y_i, \hat{p}^{(t-1)}(x_i))$ denote the negative log-likelihood for observation i (y_i, x_i) at step $t-1$ (this function is also called cross-entropy in the learning literature). At step t , the algorithm tries

to find π_t that minimizes

$$\sum_{i=1}^n \partial_2 \ell(y_i, \hat{p}^{(t-1)}(x_i)) \times \pi_t(x_i) + \frac{1}{2} \partial_2^2 \ell(y_i, \hat{p}^{(t-1)}(x_i)) \times \pi_t^2(x_i) + \text{pen}(\pi_t),$$

where pen denotes the regularization penalty, and ∂_2 (resp. ∂_2^2) denotes the (resp. second order) partial derivative of a function with respect to its second argument.

4 Prediction results for the SILECC database

This section presents the prediction results obtained on the database SILECC for the different models described in Section 3. The evaluation of the performance is made through the F_1 -score, Precision and Recall curves and confusion matrices, described in Section 4.1. The results of the different models are then shown in Section 4.2.

4.1 Evaluation of the performance

To assess the performance of the different models, we have randomly split our database into a train set (80%) and test set (20%). Recall that our database is very imbalanced in the sense that the proportion of cities that have had a claim is very small.

Common methods to assess the performance of binary classifiers include true positive and true negative rates, and ROC (Receiver Operating Characteristics) curves, which display the true positive rate against the false positive rate. These methods, however, are uninformative when the classes are severely imbalanced. In this context, F_1 -score and Precision-Recall curves (PRC) have been shown to be more informative [15,28]. They are both based on the values of

$$\text{Precision}(p_c) = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

against the values of

$$\text{Recall}(p_c) = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

where p_c is a cut-off probability varying between 0 and 1. Precision quantifies the number of correct positive predictions out all positive predictions made; and Recall (often also called Sensitivity) quantifies the number of correct positive predictions out of all positive predictions that could have been made. Both focus on the Positives class (the minority class, cities with a claim) and are unconcerned with the Negatives (the majority class, cities without a claim).

The F_1 -score combines these two measures in a single index by taking the harmonic mean of those two values. It is derived from the F-Measure introduced in [18] and [27] and defined as

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (2)$$

To closer the F_1 -score is to 1, the better the prediction model is.

PRC display the values of Precision and Recall as the cut-off probability p_c varies from 0 to 1. The PR curve of a skillful model bows towards the point with coordinates (1, 1). The curve of a no-skill classifier will be a horizontal line on the plot with a y-coordinate proportional to the number of Positives in the dataset. For a balanced dataset this will be 0.5 [15].

PRC and F_1 -score are complementary in our approach, the PRC is used on the probability outcomes of the models, it gives the best configuration and the best model. Whereas the F_1 -score is used to select the best threshold value used in the prediction of the two classes, for each model. We will detail this process in the next section.

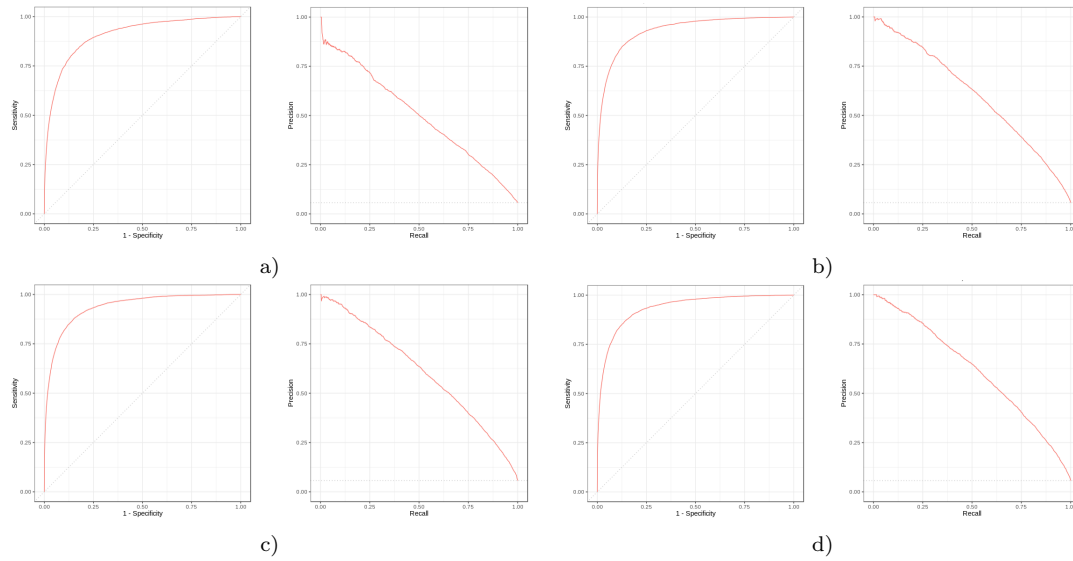


Fig. 6 ROC (Left) and PR (Right) curves for a) GLMNET b) RF c) XGBOOST d) AGGREGATE on the test sample (containing 5 924 cities with a claim and 98 596 without a claim).

4.2 Results

In this section, we present the main results of our analysis for the models presented in Section 3. We also consider the aggregation of these 3 models. For the aggregation (AGGREGATE), we examined the mean of the three probability outcomes of each model, to make a synthesis of all of the predictions. For each, Figure 6 show ROC and PRC and Table 1 the Area Under the Curve (AUC) for both ROC and PRC. The closer the AUC is to 1 the better is the prediction method.

MODEL	AUC ROC	AUC PRC
GLMNET	0.907	0.503
RF	0.933	0.604
XGBOOST	0.936	0.609
AGGREGATE	0.936	0.615

Table 1 AUC of the different models

In Table 1, notice that the AUC for the ROC are close to 1 for all four models while AUC for the PRC have values around 0.60. This illustrates what has been explained in Section 4.1, ROC curves are uninformative for imbalanced dataset, since ROC curves focus equally on the Positives and Negatives classes. Therefore, when the Negatives class is largely predominant in the dataset, a model always predicting a Negative will have an AUC close to 1, but will not predict any of the Positives.

The AUC PRC in Table 1 show that the XGBOOST and RF seem to provide better results on our data than the GLMNET. However, the aggregation gives the best results which is a strong advocate for the use of this method. The PR curves in Figure 6 seem to be reasonable given our classification problem, they are not perfect but the trade-off is acceptable. The XGBOOST and RF seem to work better, as the beginning of the graph decreases slower in comparison to the

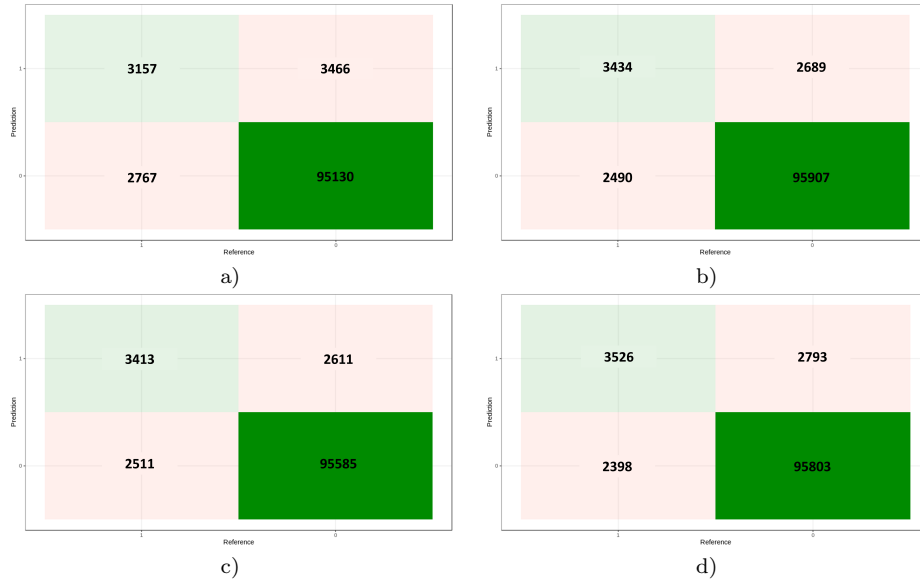


Fig. 7 Confusion matrices for a) GLMNET b) RF c) XGBOOST d) AGGREGATE on the test sample.

one in the GLMNET. It is also very interesting to see that the aggregation smooths the results and takes the best out of each model. The beginning of the graph is less discontinuous. The aggregation appears to be the best model to use for the prediction.

We then selected a threshold to make the prediction, that is the value of the cut-off probability p_c over which we consider the prediction to be a 1. A classical value of 0.5 seems arbitrary as the predictions we try to make are rare and so a probability of 0.5 can already be a strong score. To do that we used the F_1 -score and tried different thresholds with a step of 0.001. The results are presented in table 2.

MODEL	F_1 -score	Thresholds
GLMNET	0.503	0.221
RF	0.570	0.306
XGBOOST	0.573	0.291
AGGREGATE	0.576	0.264

Table 2 Best F_1 -score and thresholds associated for each model.

Again we can see that the best F_1 -score is obtained by the aggregation of the model, with a value of 0.576 and a threshold of 0.264. We find the same order as before, the XGBOOST works best than the Random Forest. The thresholds ranges between 0.2 and 0.3 which confirms the idea that the threshold 0.5 would have been arbitrary and would not make the best prediction.

Figure 7 shows the confusion matrices for each model. Out of the 5 924 cities affected by a claim in the test set, the all four models predicted between 3 157 and 3 526 cities affected by a claim. This means that they manage to predict more than half of them however there are some false predictions. GLMNET does more false predictions, which can explain the differences seen in the F_1 -score. When we examine the confusion matrices there are no notable differences between the other models.

4.3 Variable importance

Variable importance measures how much a variable influences the predictions made from a given model. The more a model relies on a variable to make predictions, the more important it is for the model. Variable importance provides also a tool to interpret “black-box” models and their performance [29]. The variable importance is based on different metrics depending on the considered model. For GLMNET, it is measured by the value of the coefficient associated with the variable, after standardization of the data. Concerning RF, the variable importance corresponds to the Gini index for classification [29,33]. For XGBOOST, the variable importance represents the fractional contribution of each feature to the model based on the total gain of the splits of the feature. A higher percentage means a more important predictive feature. Table 3 reports the top 10 contributing variables according to the relevant metrics for each model.

Table 3 Top 10 of importance variable according to the relevant metrics for each model

GLMNET	RF	XGBOOST
Max value of the SSWI 12 for February	Number of past declaration of natural catastrophe	Number of past declaration of natural catastrophe
Max value of the SSWI 12 for August	Surface with no propensity of shrinking and swelling of clay	Number of events for the previous year
Max value of the SSWI 6 for November	Number of houses	Number of houses
Max value of the SSWI 12 for June	Proportion of the surface with weak propensity of shrinking and swelling of clay	Surface with weak propensity of shrinking and swelling of clay
Max value of the SSWI 3 for August	Surface with urban area	Min value of the SSWI 1 for August
Ranking of the severity of the events	Surface with medium propensity of shrinking and swelling of clay	Min value of the SSWI 3 for October
Max value of the SSWI 12 for June	Number of houses with medium propensity of shrinking and swelling of clay	Number of past refusal with the calculation done with our SSWI
Min value of the SSWI 12 for June	Total Duration of the events of drought	Surface with medium propensity of shrinking and swelling of clay
Max value of the SSWI 12 for January	Number of events for the previous year	Number of houses with medium propensity of shrinking and swelling of clay
Max value of the SSWI 6 for January	Min value of the SSWI 6 for November	Total Duration of the past events of drought

GLMNET seems to rely more on meteorological data and less on the description of the city. Both tree methods, RF and XGBOOST rely more on a combination of exposition to shrinking and swelling of clay and past declaration of events than of the meteorological data. We only show the top 10 and the meteorological data are still important for RF and XGBOOST, they come just after. Nevertheless it is interesting to note that they are not the most important variables. This could explain the differences we saw in the results, especially as RF and XGBOOST seem to give very similar results, in terms of AUC PRC score but also for the number of cities correctly predicted. Moreover, many of available variables are not used, which is expected as we choose to give a lot of covariates and use methods that will be able to choose the relevant ones.

5 Cost prediction

In this section, we explain our methodology to predict the cost of a drought event. The model for evaluating the cost is described in Section 5.1 and the final results are shown in Section 5.2

5.1 Linear regression

We already mentioned that the cost of a drought event is likely to be correlated with the number of houses in the area. Based on the database SILECC, we fitted a model to quantify this impact. Since the database represents 70% of the market, we multiplied each cost by 1.42 to obtain a cost for the entire French market. We aggregated several events of a same year to reduce the variability in the estimation. The largest yearly cost was observed in 2003 with 2 billion euros and the average cost between 2003 and 2017 was 415 million euros per year. The number of houses has a similar distribution, going up to 4,7 millions in 2003, with an average of 1.7 million houses per year. Denoting M as the cost of an event, we can write

$$\mathbb{E}[M] = \text{Number of houses} \times 464.4 - 4.121e + 08, \quad (3)$$

with standard errors of $1.15e + 08$ for the intercept and 55.79 for the number of houses. We found a good correlation between these two variables in our database, with a $R^2 = 0.84$ and a residual standard error of $2.198e + 08$. Figure 8 shows that the linear regression is a good approximation for our problem. We use the 95% confidence interval, which we will use for our prediction.

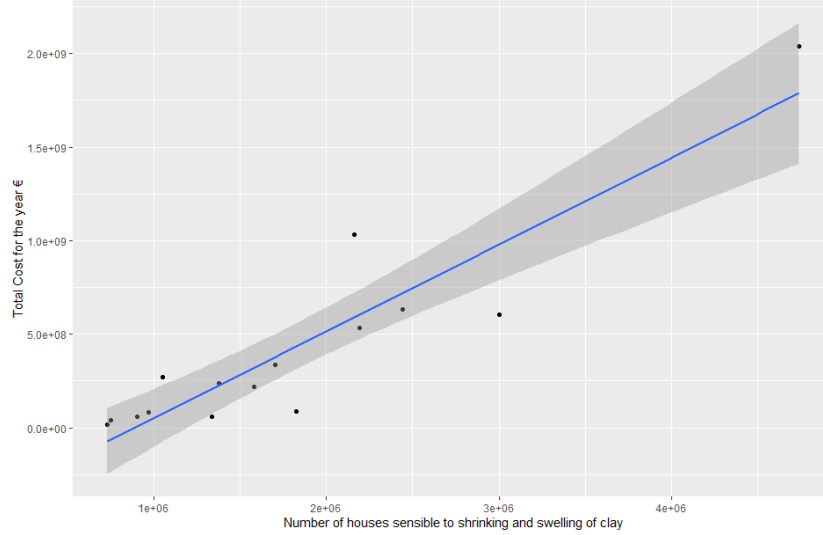


Fig. 8 Linear regression for the cost of claims Vs. number of houses. The points are the observations, the blue line the regression line and the grey are the confidence band.

This linear model is of course very rough, but has to be fitted on the small number of observations that we have (only 15), which explains the choice for the most simple regression model. Although the R^2 is relatively close to 1, one should of course not be overconfident on this fit due to the small number of points used to estimate the parameters of the model.

5.2 Results of the cost prediction for 2018

The previous model is then linked with the prediction models of Section 3. Once we have the number of houses we can estimate a cost with a confidence interval. In our case, the total loss can be written as

$$L = \sum_{i=1}^N Y_i M_i,$$

where $Y_i = 1$ if a claim occurred and 0 otherwise, and M_i is the corresponding amount of the claim (for which we know the number of houses n_i), and N is the number of considered cities. If Y_i and M_i are independent, then the variance is

$$\text{Var}(Y_i M_i | X) = (\mathbb{E}[M^2] p_i (1 - p_i) + p_i \text{Var}(M)),$$

where $p_i = \mathbb{P}(Y_i = 1|X)$. Hence the variance σ^2 of L can be estimated by

$$\hat{\sigma}^2 = \sum_{i=1}^n (\hat{m}_{2,i} p_i (1 - p_i) + p_i \tilde{\sigma}^2),$$

where $\tilde{\sigma}$ is the estimated standard error in the linear regression model of Section 5.1, and

$$\hat{m}_{2,i} = \tilde{\sigma}^2 + (\hat{\alpha} + \hat{\beta} \mathbf{n}_i)^2,$$

with $(\hat{\alpha}, \hat{\beta}) = (-4.121e + 08, 464.4)$, as estimated in the previous section.

Then the 95% confidence intervals of our estimation can be approximated by

$$\hat{L} \pm 1.96 \hat{\sigma}^2$$

The results of this estimation are displayed in Table 4. The FFA estimated the cost of the drought in France for 2018 to 900 million euros [10], with competing estimations between 1 100 and 1 300 million euros [1]. The outcome of the aggregation model shows the same results. Even if we do not have a very good precision at the city level, the general cost is consistent with the observed data.

MODEL	Lower bound	Estimate	Upper bound
GLMNET	461 125 885	579 350 811	697 575 737
RF	1 396 432 680	1 618 225 685	1 840 018 69
XGBOOST	839 262 189	977 086 655	1 114 911 122
AGGREGATE	796 820 728	965 750 651	1 134 680 547

Table 4 Estimates and confidence intervals of the predicted costs for 2018 (in euros)

Among these four classes of predictors, the penalized Generalized Linear Model tends to considerably underestimate the cost, being significantly lower than the benchmark evaluation of the FFA. The three other methods provide more plausible results, probably due to the fact that they are more flexible than a simple parametric approach, and therefore have more ability to capture complex phenomena. On the other hand, the random forests produce an estimation even beyond the most pessimistic evaluations of the risk by the market, which seems to advocate for relying on the estimations of the two other approaches.

The difference in the predicted costs is essentially due to a difference in the predicted number of houses. Which, itself, is related to the number of cities predicted by each model, as we can notice in Table 5. We can note for 2018 that RF predict more cities with a claim whereas GLMNET predicts less. This once again advocates for the use of the aggregation because it has an averaging effect on the prediction and allows us to take the best of each prediction.

MODEL	Number of cities	Number of sensitive houses
GLMNET	1 364	2 134 000
RF	5 525	4 371 000
XGBOOST	1 800	2 991 000
AGGREGATE	1 823	2 966 000

Table 5 Predicted number of sensitive houses and cities for the 2018-drought

6 Conclusion and discussions

In this work, we developed a methodology to predict the cost of the consequences of drought for the entire French market. We first used a Generalized Linear Model with Elastic-Net penalization, Random Forest and Extreme Gradient Boosting models with different discriminant thresholds to predict the cities that may be affected by a claim. Based on these predictions, we calculated the number of houses that have a propensity to clay shrinking and swelling and then computed the total cost through a linear regression.

We obtained encouraging results for such a complex phenomenon, although several uncertainties remain. Despite moderate results for the prediction of the impacted cities, we obtained coherent results for the cost prediction. The database we used, the process of natural catastrophes and the nature of this hazard of drought make the modeling very complex and uncertain. Indeed, our database is based on past claims, reported by different insurers and contain some imprecision, that may impact the results of the prediction of the probability for a city to be affected by a claim.

The second difficulty is the process of decrees of natural disasters. To be able to obtain compensation, and therefore to appear in our database, the city must have been recognized by the Commission. There may be claims in cities that were not recognized. Our models might predict such claims, but we are not able to assess whether the prediction are correct or not. Also, in the past 20 years, the criteria for a city to be recognized in the state of natural catastrophe has changed six times. Therefore, in our train database, we may have different characteristics that will have different effects depending on the criterion.

Moreover, with the meteorological and geological variables that were at our disposal, we only addressed a part of the factors that drive the risk. The interaction between the structure of the house and the composition of the soil plays an important role to determine whether the house will be damaged by a drought event. We took the nature of the soil into account with the BRGM indicators but it is very difficult to take the structure of the house into account due to the lack of data on the different types of foundations, especially at a local level.

We also faced difficulties to assess our model. As mentioned above there is uncertainty on the results due to the recognition process. More generally it is difficult to find the right score to judge a model, especially with imbalanced data. Furthermore, the prediction that we make can only be verified for one or two years, and even more if we want to have all the claims. The results are encouraging but need to be consolidated by more accurate predictions.

Despite these difficulties, the methods we developed allowed us to improve the prediction of the costs from drought in this particular French context. The techniques we used could be improved with additional amount of data, and with additional knowledge on the spatial dependence phenomena between cities (namely how two close cities may coordinate or not their responses). Let us point that the main advantage of our approach is to provide a fast answer to the question of the cost of such natural events, in a context where the time to react is important to optimize risk management. Finally, let us mention that the methods we developed could also be extended to approximate or predict the index used by the CatNat Commission, in order to improve the prediction. Indeed, this index has recently been made available.

Declarations

Funding This research was supported by the Mission Risques Naturels.

Conflict of interest The authors declare that they have no conflict of interest.

Availability of data and material The database is not publicly available for confidential reasons.

Code availability The code is publicly available at : https://github.com/antoine-heranval/Paper_Application-of-machine-learning-methods-for-cost-prediction-of-drought-in-France

References

1. Catastrophes naturelles : la facture salée des sécheresses à répétition. Tech. rep., Argus de l'Assurance. URL <https://www.argusdelassurance.com/assurance-dommages/catastrophes-naturelles-la-facture-salee-des-secheresses-a-repetition.169969>
2. Avant de construire – prendre en compte les risques du terrain. Tech. rep., Agence Qualité Construction (2014). URL <https://qualiteconstruction.com/publication/avant-de-construire-prendre-en-compte-les-risques-du-terrain/>
3. Etude : Changement climatique et assurance à l'horizon 2040. Tech. rep., Fédération Française de l'assurance (2015). URL <https://www.ffa-assurance.fr/la-federation/publications/enjeux-climatiques/etude-changement-climatique-et-assurance-horizon-2040>
4. Présentation de la MRN. Tech. rep., Mission Risques Naturels (2018). URL https://www.mrn.asso.fr/wp-content/uploads/2018/09/presentation-mrn_v21092018-1.pdf
5. Sécheresse géotechnique, de la connaissance de l'aléa à l'analyse de l'endommagement du bâti. Tech. rep., Mission Risques Naturels (2018). URL https://www.mrn.asso.fr/wp-content/uploads/2019/01/21-01-2018_rapport-mrn_secheresse-2018.pdf
6. Lettre d'information de la Mission Risques Naturels 30, juillet 2019. Tech. rep., Mission Risques Naturels (2019). URL https://www.mrn.asso.fr/wp-content/uploads/2019/10/lettre-n30_vf.pdf
7. Procédure de reconnaissance de l'état de catastrophe naturelle - Révision des critères permettant de caractériser l'intensité des épisodes de sécheresse-réhydratation des sols à l'origine de mouvements de terrain différentiels. Tech. rep., Ministère de l'intérieur (2019). URL <https://www.legifrance.gouv.fr/download/pdf/circ?id=44648>
8. Contribution de Météo-France à l'analyse de la sécheresse géotechnique à l'attention de la Commission CatNat pour l'année 2019. Tech. rep., Météo France, Direction de la Climatologie et des Services Climatiques (2020). URL <http://www.meteofrance.fr/documents/10192/36885873/Rapport-CatNat-Secheresse-2020.pdf>
9. Météo-France dans le dispositif CATNAT sécheresse. Tech. rep., Météo France (2020). URL <http://www.meteofrance.fr/documents/10192/79826318/Meteo-France+dans+le+dispositif+CATNAT+secheresse>
10. L'assurance des événements naturels en 2019. Tech. rep., Fédération Française de l'assurance (2021). URL <https://www.mrn.asso.fr/wp-content/uploads/2021/03/2021-mrn-lassurance-des-evenements-naturels-en-2019.pdf>
11. Arnold, C.: Le parc de logements en France au 1er janvier 2018. Tech. rep., INSEE (2018). URL <https://www.insee.fr/fr/statistiques/3620894>
12. Assadollahi, H.: The impact of climatic events and drought on the shrinkage and swelling phenomenon of clayey soils interacting with constructions. Ph.D. thesis, Université de Strasbourg (2019). URL https://tel.archives-ouvertes.fr/tel-02331567/file/Assadollahi_Hossein_2019_ED269.pdf
13. Breiman, L.: Random forests. *Machine Learning* **45**, 5–32 (2001). DOI 10.1023/A:1010933404324
14. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. CRC press (1984). DOI 10.1201/9781315139470
15. Brownlee, J.: Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. *Machine Learning Mastery* (2020)
16. Charpentier, A., James, M.R., Ali, H.: Predicting drought and subsidence risks in France. *Natural Hazards and Earth System Sciences Discussions* pp. 1–27 (2021)
17. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM. DOI 10.1145/2939672.2939785
18. Chinchor, N., Sundheim, B.M.: Muc-5 evaluation metrics. In: *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993* (1993). URL <https://aclanthology.org/M93-1007.pdf>
19. Denuit, M., Charpentier, A.: *Mathématiques de l'Assurance Non-Vie. Tome II: Tarification et Provisionnement* (2005)
20. Ecoto, G., Bibaut, A., Chambaz, A.: One-step ahead sequential super learning from short times series of many slightly dependent data, and anticipating the cost of natural disasters. arXiv preprint arXiv:2107.13291 (2021)
21. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**(1), 1 (2010). URL <https://pubmed.ncbi.nlm.nih.gov/20808728/>

22. Habets, F., Boone, A., Champeaux, J.L., Etchevers, P., Franchistéguy, L., Leblois, E., Ledoux, E., Le Moigne, P., Martin, E., Morel, S., Noilhan, J., Quintana Seguí, P., Rousset-Regimbeau, F., Viennot, P.: The SAFRAN-ISBA-MODCOU hydrometeorological model applied over France **113**, D06113. DOI 10.1029/2007JD008548
23. Marquardt, D.W., Snee, R.D.: Ridge regression in practice. *The American Statistician* **29**(1), 3–20 (1975). DOI 10.1080/00031305.1975.10479105
24. McKee, T.B., Doesken, N.J., Kleist, J., et al.: The relationship of drought frequency and duration to time scales. In: *Proceedings of the 8th Conference on Applied Climatology*, vol. 17, pp. 179–183. Boston (1993). URL <https://climate.colostate.edu/pdfs/relationshipofdroughtfrequency.pdf>
25. Nelder, J.A., Wedderburn, R.W.: Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135**(3), 370–384 (1972). DOI 10.2307/2344614
26. Pritchard, O.G., Hallett, S.H., Farewell, T.S.: Probabilistic soil moisture projections to assess Great Britain’s future clay-related subsidence hazard **133**(4), 635–650. DOI 10.1007/s10584-015-1486-z
27. Rijsbergen, C.: *Information retrieval 2nd ed buttersworth*. London (1979)
28. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets **10**(3), e0118432. DOI 10.1371/journal.pone.0118432
29. Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288 (1996). DOI 10.1111/j.2517-6161.1996.tb02080.x
30. Vidal, J.P., Martin, E., Kitova, N., Najac, J., Soubeyroux, J.M.: Evolution of spatio-temporal drought characteristics: validation, projections and effect of adaptation scenarios **16**(8), 2935–2955. DOI 10.5194/hess-16-2935-2012
31. Vidal, J.P., Moisselin, J.M.: *Impact du changement climatique sur les sécheresses en France* (2011). URL http://www.drias-climat.fr/public/shared/rapport_final_CLIMSEC.pdf
32. Vincent, M., Plat, E., Le Roy, S.: Cartographie de l’aléa retrait-gonflement et plans de prévention des risques. *Revue française de géotechnique* (120-121), 189–200 (2007). DOI 10.1051/geotech/2007120189. URL <https://www.georisques.gouv.fr/articles-risques/exposition-du-territoire-au-phenomene>
33. Wright, M.N., Ziegler, A.: ranger: A fast implementation of random forests for high dimensional data in C++ and R **77**(1). DOI 10.18637/jss.v077.i01
34. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (statistical methodology)* **67**(2), 301–320 (2005). DOI 10.1111/j.1467-9868.2005.00503.x