



HAL
open science

Modeling and machine learning applied to the assessment of the cost of a drought event within the French natural disaster compensation scheme

Geoffrey Ecoto Dicka

► **To cite this version:**

Geoffrey Ecoto Dicka. Modeling and machine learning applied to the assessment of the cost of a drought event within the French natural disaster compensation scheme. Statistics [math.ST]. Université Paris Cité, 2023. English. NNT : 2023UNIP7182 . tel-04637025

HAL Id: tel-04637025

<https://theses.hal.science/tel-04637025v1>

Submitted on 5 Jul 2024

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.

UNIVERSITÉ PARIS CITÉ

ÉCOLE DOCTORALE
Sciences Mathématiques de Paris Centre (386)
Laboratoire MAP5, CNRS UMR 8145

THÈSE DE DOCTORAT

présentée par :

Geoffrey ECOTO DICKA

En vue de l'obtention du grade de **docteur de Université Paris Cité**
Spécialité : MATHÉMATIQUES APPLIQUÉES

**Modélisation et apprentissage machine learning appliqués à
l'estimation des dommages consécutifs à la survenance d'un
événement de sécheresse par retrait-gonflement des argiles
dans le cadre du régime d'indemnisation des catastrophes
naturelles français**

Soutenue le 19 décembre 2023

Directeur de thèse : Antoine CHAMBAZ

JURY :

Jean ARDON, PhD	MAIF	Examinateur
Laurence BARRY, PhD	Chaire PARI, ENSAE/Sciences Po	Examinatrice
Antoine CHAMBAZ, PR	Université Paris Cité	Directeur
Arthur CHARPENTIER, PR	Université du Québec à Montréal, Université de Rennes	Rapporteur
Marianne CLAUSEL, PR	Université de Lorraine	Examinatrice
Thierry COHIGNAC, PhD	CCR	Co-encadrant
Caroline HILLAIRET, PR	ENSAE	Examinatrice
Olivier LOPEZ, PR	ENSAE	Rapporteur

UNIVERSITÉ PARIS CITÉ

ÉCOLE DOCTORALE
Sciences Mathématiques de Paris Centre (386)
Laboratoire MAP5, CNRS UMR 8145

THÈSE DE DOCTORAT

présentée par :

Geoffrey ECOTO DICKA

En vue de l'obtention du grade de **docteur de Université Paris Cité**

Spécialité : MATHÉMATIQUES APPLIQUÉES

**Modélisation et apprentissage machine learning appliqués à
l'estimation des dommages consécutifs à la survenance d'un
événement de sécheresse par retrait-gonflement des argiles
dans le cadre du régime d'indemnisation des catastrophes
naturelles français**

Soutenue le 19 décembre 2023

Directeur de thèse : Antoine CHAMBAZ

JURY :

Jean ARDON, PhD	MAIF	Examinateur
Laurence BARRY, PhD	Chaire PARI, ENSAE/Sciences Po	Examinatrice
Antoine CHAMBAZ, PR	Université Paris Cité	Directeur
Arthur CHARPENTIER, PR	Université du Québec à Montréal, Université de Rennes	Rapporteur
Marianne CLAUSEL, PR	Université de Lorraine	Examinatrice
Thierry COHIGNAC, PhD	CCR	Co-encadrant
Caroline HILLAIRET, PR	ENSAE	Examinatrice
Olivier LOPEZ, PR	ENSAE	Rapporteur

Résumé

Cette thèse est consacrée à l'anticipation de l'impact financier sur les biens assurés de la survenance d'un événement de sécheresse grâce au recours à des méthodes au croisement de la statistique et du machine learning. Le terme sécheresse désigne ici le phénomène de retrait-gonflement des argiles provoquant des dommages aux bâtiments. L'exercice peut être décomposé en deux sous-problèmes que nous abordons tour à tour. Le premier sous-problème considère plus spécifiquement la tâche consistant à prédire quelles communes formuleront une demande de reconnaissance de l'état de catastrophe naturelle au titre de l'événement sécheresse. Le second est consacré à la prédiction de l'impact financier de l'événement sécheresse sur les biens assurés situés dans les communes reconnues en état de catastrophe naturelle. Dans le cadre du premier sous-problème, nous développons, étudions et appliquons un algorithme original pour la prédiction des demandes de reconnaissance de l'état de catastrophe naturelle. L'algorithme bénéficie de deux formalisations complémentaires de la tâche d'intérêt, abordé sous l'angle de la classification supervisée et comme un problème de transport optimal. Les prédictions finales sont obtenues comme moyenne géométrique des deux types de prédictions. Théoriquement, le plan de transport optimal peut être obtenu en appliquant l'algorithme iPiano [Ochs et al., 2015], dont nous prouvons que les hypothèses qui sous-tendent son analyse sont bien vérifiées. L'analyse des prédictions obtenues démontre la pertinence de l'algorithme. Dans le cadre du second sous-problème, nous développons, étudions et appliquons un algorithme original d'agrégation d'algorithmes inspiré du Super Learner [van der Laan, 2007]. Deux écueils doivent être pris en compte. D'une part, parce que le péril sécheresse n'est couvert par le régime d'indemnisation des catastrophes naturelles français que depuis 1989, le nombre d'événements sécheresse sur lesquels nous pouvons entraîner notre algorithme est réduit, chaque événement sécheresse se voyant associer un jeu de données de grande taille. D'autre part, à la dépendance temporelle s'ajoute une dépendance spatiale due notamment aux proximités géographique et administrative entre communes françaises. Fondée sur une modélisation de la dépendance à l'aide d'un graphe de dépendance, l'étude théorique révèle que la brièveté de la série temporelle peut être compensée si la dépendance spatiale est faible. De nouveau, l'analyse des prédictions obtenues démontre la pertinence de notre algorithme.

Mots-clés : catastrophes naturelles, machine learning, événements sécheresse, statistique, super learning, transport optimal.

Abstract

This Ph.D. thesis is dedicated to forecasting the financial impact on insured properties in the event of drought, utilizing methods that merge statistics and machine learning. In this context, "drought" refers to the phenomenon of clay shrinkage and swelling that leads to damage to buildings. The task can be broken down into two sub-problems that we address separately. The first sub-problem focuses on predicting which municipalities will submit a request for the government declaration of natural disaster for a drought event. The second is dedicated to predicting the financial impact of drought events on insured properties located in municipalities that obtained the government declaration of natural disaster for a drought event. For the first sub-problem, we develop, study, and apply an original algorithm to predict requests for the government declaration of natural disaster. The algorithm benefits from two complementary formalizations of the task at hand, approached as both supervised classification and an optimal transport problem. The final predictions are obtained as a geometric mean of these two prediction types. Theoretically, the optimal transport plan can be obtained by applying the iPi-ano algorithm [Ochs et al., 2015], and we demonstrate that the assumptions underlying its analysis are met. The analysis of the predictions obtained confirms the algorithm's relevance. Regarding the second sub-problem, we develop, investigate, and apply an original aggregation algorithm, inspired by the Super Learner [van der Laan, 2007]. Two challenges must be considered. First, since drought events have only been covered by the French natural disaster compensation scheme since 1989, the number of drought events available for training our algorithm is limited, with each drought event associated with a large dataset. Second, temporal dependence is compounded by spatial dependence, primarily due to geographic and administrative proximity between French municipalities. Based on a dependency modeling using a dependency graph, the theoretical analysis reveals that the brevity of the time series can be compensated if spatial dependence is weak. Once again, the analysis of the predictions obtained underscores the relevance of our algorithm.

Keywords: drought events, machine learning, natural disasters, optimal transport, statistics, super learning.

Table des matières

1	Introduction générale	1
1.1	Généralités	2
1.1.1	Le régime d’indemnisation des catastrophes naturelles français et la Caisse Centrale de Réassurance	2
1.1.2	Péril sécheresse : définition en enjeux financiers	3
1.1.3	De la nécessité d’anticiper les dommages	7
1.1.4	Objectifs de la thèse	8
1.2	Super Learner	9
1.2.1	L’agrégation de modèles et le Super Learner	9
1.2.2	Le Super Learner discret	10
1.2.3	Le Super Learner continu	12
1.2.4	Le Super Learner séquentiel	14
1.3	Apports de la thèse	15
1.3.1	Forecasting the cost of drought events in France by Super Learning from a short time series of many slightly dependent data	15
1.3.2	L’anticipation des communes demanderesses de la reconnaissance de l’état de catastrophe naturelle par Super Learning	16
1.3.3	Making sparse predictions, and forecasting the requests of the government declaration of natural disaster for a drought event in France	16
1.3.4	Conclusion et perspectives	16
2	Forecasting the cost of drought events in France by Super Learning from a short time series of many slightly dependent data	17
2.1	Introduction	17
2.2	Data	20
2.2.1	Data provided by CCR’s cedents	20
2.2.2	Data garnered from other sources	20
2.2.3	City-level data processing	21
2.3	The One-Step Ahead Sequential Super Learner	25
2.3.1	Aggregation strategies	25
2.3.2	Presentation and theoretical performance of an OSASSL built to forecast the cost of drought events	26

2.3.3	Forecasting the cost of drought events	29
2.4	Application	30
2.4.1	Implementing two OSASSLs	30
2.4.2	Training the discrete and continuous overarching Super Learners	32
2.4.3	Results	33
2.4.4	On the importance of the variables used to make predictions	36
2.5	Discussion	40
2.6	Appendix: the OSASSL and its oracular performances	41
2.7	Appendix: proofs	47
2.7.1	Proof of Theorem 1	47
2.7.2	Proof of Corollary 2	51
2.7.3	Proof of Theorem 3	53
2.8	Appendix: a classical strong convexity argument	56
3	L'anticipation des communes demandereses de la reconnaissance de l'état de catastrophe naturelle par Super Learning	59
3.1	Éléments sur la cinétique du péril sécheresse et estimation des demandes de reconnaissance	60
3.1.1	La sécheresse : un péril à déroulement long	60
3.1.2	Chronologie des méthodes d'estimation du coût d'un événement sécheresse	61
3.1.3	Les enjeux de l'anticipation des communes demandereses dans le cadre du provisionnement	62
3.2	Les données	63
3.3	L'OSASSL pour l'estimation des demandes de reconnaissance	65
3.3.1	L'architecture retenue et l'implémentation	66
3.3.2	Les résultats	68
3.4	Un nouvel angle de modélisation pour l'anticipation des demandes de reconnaissance	69
3.4.1	L'estimation dynamique des demandes de reconnaissance	70
3.4.2	OSASSL et transfer learning	74
3.4.3	Mise en œuvre	77
3.4.4	Les résultats	79
3.5	Éléments conclusifs	83
3.6	Perspectives de recherche	84
3.7	Annexe	86
3.7.1	Annexe A: représentations des architectures des OSASSL	86
3.7.2	Annexe B: prédictions des probabilités de demande de reconnaissance au titre de l'événement sécheresse de 2020 réalisées par le meta-algorithme fondé sur un réseau de neurones profond	87

4	Making sparse predictions, and forecasting the requests of the government declaration of natural disaster for a drought event in France	89
4.1	Introduction	90
4.2	Data and statistical challenge	91
4.2.1	Presentation of the data, first pass	91
4.2.2	Presentation of the data, second pass	92
4.2.3	The statistical challenge and some facts about the data	95
4.3	A modicum of optimal transport theory	97
4.4	Making sparse predictions	98
4.4.1	Translation to an optimization problem	98
4.4.2	On solving (4.3)	99
4.4.3	Implementation of the “OT-procedure”	101
4.5	A simple simulation study, introducing the “hybrid procedure”	102
4.5.1	Simulated data	102
4.5.2	Fine-tuning the OT-procedure	102
4.5.3	Alternative, classification-based approaches	103
4.5.4	Results, introducing the “hybrid procedure”	103
4.6	Forecasting the requests of the government declaration of natural disaster for a drought event in France	104
4.6.1	Fine-tuning the OT-procedure	104
4.6.2	Alternative, classification-based approaches	111
4.6.3	Results	112
4.6.4	On the importance of the variables used to make predictions	118
4.7	Discussion	121
4.8	Appendix: checking the iPiano assumptions	125
4.8.1	The function \mathbf{f} is C^1 -smooth and its gradient is Lipschitz continuous on $\mathbf{dom} \mathbf{g}_\tau$	125
4.8.2	The function \mathbf{H}_δ satisfies the Kurdyka-Lojasiewicz property	131
5	Conclusion et perspectives	137
	Bibliographie	139

Table des figures

1.1	Le phénomène de retrait-gonflement des argiles provoque des fissures sur les maisons. La présence d’arbres à proximité du bâtiment est un facteur aggravant du fait de l’aspiration par les racines de l’eau du sol.	3
1.2	La sinistralité sécheresse en France de 1989 à 2022 (source CCR). Les montants affichés sont en millions d’euros et actualisés en euros 2022. Ces chiffres correspondent à l’ensemble du marché français et comprennent donc une part de sinistralité non couverte par CCR.	5
1.3	Cartographie de l’exposition du territoire au phénomène de retrait-gonflement : 48% du territoire est en zone d’exposition moyenne ou forte (source BRGM). Quatre catégories sont présentées et caractérisent l’exposition au RGA : à priori nulle en blanc, faible en jaune, moyenne en orange et forte en rouge.	6
1.4	Les 3 modules d’un modèle catastrophe : aléa, vulnérabilité et dommages	7
2.1	Estimated overall costs of drought events across France (blue) and provisional city-specific costs obtained by aggregating the costs of those claims filled in the claims data provided by the cedents (yellow). The ratios of the latter to the former (numbers above the blue bars) range between 22% and 97% (the 144%-ratio corresponds to an exceptional year where the estimated overall cost is even smaller than the very small aggregated cost of the claims data). In this figure we use current euros. Source: CCR. . .	22
2.2	Chunks from five arbitrarily chosen time series of Soil Wetness Index (SWI) over the course of one year. It does not come as a surprise that the soil is drier during summer than during winter.	23
2.3	Evolution (from 2007 onward) of the weights attributed in the overarching Super Learner to four of the algorithms $\hat{\theta}_1, \dots, \hat{\theta}_J$. The others get no weight at all.	33
2.4	Presentation (from 2007 onward) of the real costs of drought events and their predictions. The predictions are either those made by the discrete (pale yellow) and continuous overarching (dark yellow) Super Learners or obtained by averaging all the base learners’ predictions (red) or using their median (blue vertical bars). The figure also presents boxplots that summarize all the base learners’ predictions. Note the high variability of these predictions. In this figure we use current euros.	36

2.5	Kernel density estimates of the conditional laws of the residual error (of the predictions made by the continuous overarching Super Learner) in ten strata characterized by the deciles of the city-level costs. The cross at the upper RHS of the plot indicates the maximum residual error, made for a city belonging to the last decile of the cost distribution. In this figure we use current euros.	37
2.6	Geographical distribution of the residual errors (of the predictions made by the continuous overarching Super Learner). (A): a city contributes as many points as the number of times it benefited from a government declaration of natural disaster for a drought event between 2007 and 2017. (B): a city contributes a point if and only if it benefited from a government declaration of natural disaster for a drought event in 2016. The color reflects the quartile of the residual error to which the city- and time-specific residual error belongs (based on all the errors). In (A), the transparency reflects the number of times the city benefited from a government declaration of natural disaster for a drought event between 2007 and 2017, a larger number leading to less transparency.	38
2.7	Assessing the importance of the variables used to make predictions by the discrete overarching Super Learner. The larger is ρ_s the more we are willing to believe that the s -th covariate well explains the predictions made by the discrete overarching Super Learner. Every ρ_s is declared significantly positive by a permutation test analysis.	39
3.1	Rythme de paiement des sinistres liés à la sécheresse RGA effectués par CCR. Sept ans et demi après la survenance d'un événement sécheresse, 75% de sinistres ont été payés en moyenne.	60
3.2	Chronologie des méthodes de prédiction déployées par CCR et évolution des communes demanderesses connues.	61
3.3	Nombre de demandes de reconnaissance de l'état de catastrophe naturelle au titre de la sécheresse RGA par année.	64
3.4	Somme des probabilités de demande prédites annuellement par l'overarching discret et continu, comparaison aux demandes réelles.	68
3.5	Cartographie des prédictions des probabilités de demande pour l'événement sécheresse de 2021 réalisées par le Super Learner overarching discret. . . .	69
3.6	Rythme de dépôt des demandes de reconnaissance de l'état de catastrophe naturelle au titre de la sécheresse RGA observé sur les événements 2019 à 2021.	70
3.7	Évolution du nombre de demandes de reconnaissance de l'état de catastrophe naturelle au titre de la sécheresse RGA observé sur les événements 2019 à 2021.	71

3.8 MSE, stock de demandes et somme des prédictions des probabilités de demande de reconnaissance au titre de l'événement sécheresse 2021 pour chaque semaine étudiée. La MSE est représentée par la courbe noire correspondant à la fonction $u \mapsto \frac{1}{|\mathcal{A}_{2021,u}^-|} \sum_{\alpha \in \mathcal{A}_{2021,u}^-} (\zeta_{\alpha,2021} - \widehat{\alpha}_{\alpha,2021,u})^2$. Le trait bleu en pointillés représente le nombre de demandes réelles. 81

3.9 Cartographie des prédictions des probabilités de demande pour l'événement sécheresse de 2021 réalisées par l'overarching Super Learner continu avec transfer learning en semaine 49. Sur la carte de gauche, les communes appartenant au stock sont en bleu. 82

3.10 Cartographie des prédictions des probabilités de demande pour l'événement sécheresse de 2021 réalisées par l'overarching Super Learner continu avec transfer learning en semaine 78. Sur la carte de gauche, les communes appartenant au stock sont en bleu. 83

3.11 Représentation de l'architecture de l'OSASSL déployé pour l'estimation des demandes de reconnaissance. 86

3.12 Représentation de l'architecture de l'OSASSL déployé pour l'estimation des demandes de reconnaissance dans le cadre de l'approche dynamique avec transfer learning. 86

3.13 MSE, stock de demandes et somme des prédictions des probabilités de demande de reconnaissance au titre de l'événement sécheresse 2020 pour chaque semaine étudiée. La MSE est représentée par la courbe correspondant à la fonction $u \mapsto \frac{1}{|\mathcal{A}_{2020,u}^-|} \sum_{\alpha \in \mathcal{A}_{2020,u}^-} (\zeta_{\alpha,2020} - \ell_{5,2019}(X_{\alpha,2020,u}))^2$. Le trait bleu en pointillés représente le nombre de demandes réelles. Les prédictions sont réalisées par un meta-algorithme fondé sur un réseau de neurones profond ($k = 5$). 87

4.1 Empirical cumulative distribution functions of the sets $\{\widehat{y}_{n,\ell}^\bullet : \ell \in \llbracket L \rrbracket, n \in \llbracket N \rrbracket \text{ st } y'_{n,\ell} = y\}$ for $y = 0$ (left-hand side panel) and $y = 1$ (right-hand side panel), where the symbol \bullet stands for $\text{SL}_1, \text{SL}_2, \text{OT}, \text{HYB}$ 105

4.2 Scatterplot of $(\text{MSE}_\ell^{\text{HYB}} - \text{MSE}_\ell^\bullet) / \text{MSE}_\ell^{\text{SL}_2}$ against $(\text{MSE}_\ell^{\text{OT}} - \text{MSE}_\ell^{\text{SL}_2}) / \text{MSE}_\ell^{\text{SL}_2}$ ($\ell \in \llbracket 30 \rrbracket$) where the symbol \bullet stands for SL_2 (blue) or OT (red). See also Table 4.2. 106

4.3 Cumulative distribution functions of the sets $\{\text{cst}_{m,n} \times a_k \|\tilde{x}_{m,[k]} - \tilde{x}'_{n,[k]}\|_2^2 : m, n \in \llbracket 128 \rrbracket\}$ ($k = 1, 2, 3, 4$) where $\tilde{x}_1, \dots, \tilde{x}_{128}$ and $\tilde{x}'_1, \dots, \tilde{x}'_{128}$ are derived from x_1, \dots, x_{128} and x'_1, \dots, x'_{128} which are independently sampled, uniformly without replacement, from $\{\xi_{\alpha,1,44} : \alpha \in \mathcal{A}\}$ and $\{\xi_{\alpha,2,48} : \alpha \in \mathcal{A}\}$, where a is selected based on the HYPERBAND algorithm, and where each $\text{cst}_{m,n}$ is such that $\text{cst}_{m,n} \times \sum_{k=1}^4 a_k \|\tilde{x}_{m,[k]} - \tilde{x}'_{n,[k]}\|_2^2 = 1$ for all $m, n \in \llbracket 128 \rrbracket$. The more a cumulative distribution function is shifted to the right the more a generic sum $\sum_{k=1}^4 a_k \|x_{[k]} - x'_{[k]}\|_2^2$ (for any $x, x' \in \mathcal{X}$, the left-hand side sum in (4.12)) is driven by the corresponding groups of covariates. See also Table 4.4. 110

- 4.4 The left-hand side maps show the probabilities predicted by the hybrid procedure of submitting a request relative to year 2021 for weeks 49 (top) and 78 (bottom). The right-hand side maps show the cities that did submit a request eventually. In both left-hand side maps, the cities for which it was already known that they submitted a request are colored in blue. It is worth emphasizing that there are no predicted probabilities within the range of 50% to 90% during week 78. 113
- 4.5 This plot shows, when week u is one of the 49th week of 2021 (December 6th to 12th), the $(59 - 52) = 7$ th, $(69 - 52) = 17$ th and $(78 - 52) = 26$ th weeks of 2022 (February 15th to 21st, April 26th to May 2nd, June 28th to July 4th), the empirical cumulative distribution functions (ecdfs) of the predicted probabilities of submitting a request made by procedures SL, OT and HYB separately for those cities that will not eventually submit a request for the government declaration of natural disaster for a drought event for year 2021 (that is, the ecdfs of $\{\widehat{\zeta}_{\alpha,3}^{\bullet,u} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3} = 0\}$, left-hand side panels) and for those that will (that is, the ecdfs of $\{\widehat{\zeta}_{\alpha,3}^{\bullet,u} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3} = 1\}$, right-hand side panels). See also Figure 4.7 for a focus on medians. 115
- 4.6 This plot shows, for week u equal either to the 49th week of 2021 (December 6th to December 12th) or the $(78 - 52) = 26$ th week of 2022 (June 27th to July 3rd), the predicted probabilities of submitting a request made by procedures SL (x -axis) and OT (y -axis) separately for those cities that will not eventually submit a request for the government declaration of natural disaster for a drought event for year 2021 (that is, $\{(\widehat{\zeta}_{\alpha,3}^{\text{SL},u}, \widehat{\zeta}_{\alpha,3}^{\text{OT},u}) : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 0\}$, left-hand side panels) and for those that will (that is, $\{(\widehat{\zeta}_{\alpha,3}^{\text{SL},u}, \widehat{\zeta}_{\alpha,3}^{\text{OT},u}) : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 1\}$, right-hand side panels). In addition, three colored points represent in each panel the coordinate-specific quantiles of order 10%, 50% and 90%. 116
- 4.7 This plot shows, as week u goes from the 49th week of 2021 (December 6th to December 12th) to the $(78 - 52) = 26$ th week of 2022 (June 27th to July 3rd), the evolutions of the medians of the predicted probabilities of submitting a request made by procedures SL, OT and HYB separately for those cities that will not eventually submit a request for the government declaration of natural disaster for a drought event for year 2021 (that is, of $u \mapsto \text{median}\{\widehat{\zeta}_{\alpha,3}^{\bullet,u} : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 0\}$, left-hand side panel) and for those that will (that is, of $u \mapsto \text{median}\{\widehat{\zeta}_{\alpha,3}^{\bullet,u} : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 1\}$, right-hand side panel). See also Figure 4.5 for more comprehensive descriptions through empirical cumulative distribution functions. 117

- 4.8 This plot shows, as week u goes from the 49th week of 2021 (December 6th to 12th) to the $(78 - 52) = 26$ th week of 2022 (June 27th to July 3rd), the evolutions of the cardinality of the stock of requests already submitted for the government declaration of natural disaster for a drought event for year 2021 (that is, of $u \mapsto \sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,3,u}$, in blue) and of the sum of the predicted probabilities that the cities which have not yet submitted such a request will eventually do (that is, of $u \mapsto \sum_{\alpha \in \mathcal{A}_t} \hat{\zeta}_{\alpha,3}^{\text{HYB},u} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$, in red). The actual eventual number of such requests (that is, $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,3}$, which equals 1696) is also represented (horizontal dashed line). In addition, the plot shows the evolution of MSE (that is, of $u \mapsto n_{3,u}^{-1} \sum_{\alpha \in \mathcal{A}_3} (\hat{\zeta}_{\alpha,3}^{\text{HYB},u} - \zeta_{\alpha,3})^2 \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ where $n_{3,u} := \sum_{\alpha \in \mathcal{A}_3} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ is the number of cities which have not submitted such a request yet at week u , in yellow). See also Table 4.5. 119
- 4.9 This plot shows, as week u goes from the 49th week of 2021 (December 6th to 12th) to the $(78 - 52) = 26$ th week of 2022 (June 27th to July 3rd), the evolutions of the importance of each variable used to make predictions, as defined in Section 4.6.4. For every eligible $s \in \llbracket 67 \rrbracket$, the larger is ρ_s^u , the stronger is the association between the s th covariate $\xi_{\alpha,3,u,s}$ and the prediction $\hat{\zeta}_{\alpha,3}^{\text{HYB},u}$ across $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. Values above the black horizontal lines are deemed highly significant based on permutation tests. See also Table 4.6. 122

Liste des tableaux

1.1	Vue synthétique des 3 grands principes d'agrégation de modèles	10
2.1	Quartiles, 99%-quantile and mean of the numbers of neighboring cities in France in 2019. Although the maximum cannot be interpreted literally as $\text{deg}(\mathcal{G}) - 1$, it nevertheless gives a sense of what a meaningful value of $\text{deg}(\mathcal{G})$ could be.	29
2.2	Averages and standard deviations (over the years) of the ratios of the predicted costs to the real costs. The predictions are those made by the discrete and continuous overarching Super Learners or derived from all the base learners' predictions by averaging, taking their median, or employing nine different aggregation strategies from the robust online aggregation literature (see main text for details).	35
3.1	Contribution des événements sécheresse aux enjeux financiers générés par les 4 derniers événements.	62
3.2	Synthèse des métriques calculées pour les prédictions du Super Learner overarching discret et continu.	68
3.3	Résumé de la distribution du minimum du SWI des communes de France métropolitaine pour les événements sécheresse 2019 et 2020, pour l'événement sécheresse 2021 et pour les événements sécheresse de 1995 à 2021.	75
3.4	Synthèse des métriques MSE et Logloss calculées pour l'événement sécheresse 2021 dans le cadre de l'approche statique, dynamique et dynamique avec transfer learning	79
4.1	Summary measures of the sets $\{\sum_{\alpha \in \mathcal{A}_t} (\zeta_{\alpha,t,u} - \zeta_{\alpha,t,u^-}) : u \in \mathcal{U}_t\}$ ($t = 1, 2, 3$), that is, of the numbers of new requests for the government declaration of natural disaster for a drought event as weeks go by, for years 2019, 2020 and 2021 respectively. In addition, the overall numbers $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t}$ and proportions $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t} / \text{card } \mathcal{A}_t$ ($t = 1, 2, 3$) of requests for the government declaration of natural disaster for a drought event relative to year t are also reported for years 2019, 2020 and 2021.	97

- 4.2 Averages and standard deviations of the mean squared errors $\{\text{MSE}_\ell^\bullet : \ell \in \llbracket L \rrbracket\}$ (4.10) where the symbol \bullet stands for $\text{SL}_1, \text{SL}_2, \text{OT}, \text{HYB}$ and $L = 30$. See also Figure 4.2. In each column, the smallest value stands out in bold characters. 107
- 4.3 Resource allocations and numbers of configurations $((r_{s,i}, n_{s,i}), i \in \{0, \dots, s\})$ in each bracket $s \in \{0, 1, 2, 3\}$ of the HYPERBAND procedure. 108
- 4.4 Quartiles of the sets $\{\|\tilde{x}_{m,[k]} - \tilde{x}'_{n,[k]}\|_2^2 : m, n \in \llbracket 128 \rrbracket\}$ ($k = 1, 2, 3, 4$) where $\tilde{x}_1, \dots, \tilde{x}_{128}$ and $\tilde{x}'_1, \dots, \tilde{x}'_{128}$ are derived from x_1, \dots, x_{128} and x'_1, \dots, x'_{128} which are independently sampled, uniformly without replacement, from $\{\xi_{\alpha,1,44} : \alpha \in \mathcal{A}\}$ and $\{\xi_{\alpha,2,48} : \alpha \in \mathcal{A}\}$. The last row recalls the four first entries of a selected based on the HYPERBAND algorithm. See also Figure 4.3. 109
- 4.5 Evolution of MSE $u \mapsto n_{3,u}^{-1} \sum_{\alpha \in \mathcal{A}_3} (\hat{\zeta}_{\alpha,3}^{\bullet,u} - \zeta_{\alpha,3})^2 \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ where $n_{3,u} := \sum_{\alpha \in \mathcal{A}_3} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ is the number of cities which have not submitted such a request yet at week $u \in \mathcal{U}_3$ and the symbol \bullet stands for $\text{SL}, \text{OT}, \text{HYB}$. In each row, the smallest value stand out in bold characters. See also Figure 4.8. 120
- 4.6 The five variables used to make predictions with the highest average importance $(\sum_{u \in \mathcal{U}_3} \rho_s^u / \text{card } \mathcal{U}_3, \text{ see definition in Section 4.6.4})$ in each group of covariates. For every eligible $s \in \llbracket 67 \rrbracket$, the larger is ρ_s^u , the stronger is the association between the sth covariate $\xi_{\alpha,3,u,s}$ and the prediction $\hat{\zeta}_{\alpha,3}^{\text{HYB},u}$ across $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. See also Figure 4.9. 123

Chapitre 1

Introduction générale

L'exposition aux catastrophes naturelles constitue un enjeu prioritaire pour l'ensemble des sociétés du fait des enjeux importants qu'elle représente, qu'ils soient environnementaux, politiques, économiques ou bien financiers. À titre d'exemple, les catastrophes naturelles survenues au cours de l'année 2022 ont engendré 275 Mds€ de dommages dans le monde [Swiss Re Institute, 2023]. En France, les catastrophes naturelles ont engendré 10 Mds€ de dommages cette même année, dont environ 3 Mds€ au titre de la sécheresse [MRN, 2023]. Cette exposition est amenée à évoluer à la hausse principalement du fait du changement climatique. Dans son étude de 2018 basée sur le scénario RCP 8.5 du GIEC¹, la Caisse Centrale de Réassurance (CCR) anticipe une hausse de la sinistralité liée aux événements naturels de 50% à horizon 2050 [CCR, 2018].

Dès 1982, l'État français s'est doté d'un régime d'indemnisation des catastrophes naturelles permettant de combler un déficit de couverture assurantielle contre les risques naturels. La loi du 13 juillet 1982 donne le cadre de ce régime, dont CCR est un acteur majeur. Cette entité propose aux assureurs qu'elle réassure² des couvertures illimitées avec la garantie de l'État. Aussi, afin d'appréhender au mieux le risque inhérent à cette activité, CCR a développé une expertise dans la modélisation des catastrophes naturelles. Ainsi, profitant des données de portefeuilles et de sinistres récoltées auprès de ses clients, CCR a débuté en 2004 l'élaboration de modèles dédiés à la quantification des dommages engendrés par la survenance d'événements naturels. Ces derniers lui permettent d'apprécier son exposition, celle de ses clients mais aussi, dans le cadre du régime d'indemnisation des catastrophes naturelles, celle de l'État français.

Cette thèse a pour objectif de proposer une approche innovante basée sur l'apprentissage statistique pour l'anticipation des dommages liés à un événement sécheresse en France métropolitaine. Dans ce premier chapitre nous précisons le contexte de l'étude ainsi que la définition et les enjeux financiers autour du péril modélisé. Le deuxième chapitre est consacré à l'estimation des dommages subis par les communes suite à un événement

1. Groupe d'experts Intergouvernemental sur l'Évolution du Climat.

2. Le terme réassurer désigne l'action d'assurer une entité étant elle-même un assureur

sécheresse. Les données exploitées sont présentées, ainsi que le schéma de validation et l’algorithme original développé et étudié dans le cadre de cette étude. Cet algorithme est aussi déployé dans le troisième chapitre pour prédire quelles sont les communes impactées par un événement sécheresse, étape préalable à l’estimation des dommages. Le quatrième chapitre présente un second algorithme pour la prédiction des communes impactées par un événement sécheresse, fondé notamment sur le transport optimal. Finalement, le dernier chapitre de cette thèse présente à la fois les conclusions des travaux réalisés et les perspectives de recherche envisagées.

1.1 Généralités

1.1.1 Le régime d’indemnisation des catastrophes naturelles français et la Caisse Centrale de Réassurance

Le régime d’indemnisation des catastrophes naturelles a été créé par la loi du 13 juillet 1982 qui impose l’inclusion d’une garantie des dommages causés par les catastrophes naturelles dans tous les contrats d’assurance dommages couvrant des biens situés en France. Ce régime repose sur deux piliers : la solidarité, fondement du mécanisme d’indemnisation, et la responsabilité. Toute indemnisation au titre de la loi de 1982 est subordonnée à trois conditions préalables qui doivent être cumulativement remplies :

- l’état de catastrophe naturelle doit avoir été constaté par un arrêté interministériel,
- les biens sinistrés doivent être couverts par un contrat d’assurance dommages aux biens,
- un lien de causalité doit exister entre la catastrophe constatée par l’arrêté et les dommages subis par l’assuré.

La reconnaissance de l’état de catastrophe naturelle est sollicitée par les communes, qui en font la demande auprès du Préfet du département. Le dossier départemental, qui peut concerner un nombre très variable de communes, est ensuite examiné par une commission interministérielle qui émet un avis sur l’état ou l’absence de catastrophe naturelle, au sens de la loi. Cette commission, qui se réunit environ une fois par mois (sauf en cas de survenance d’un événement exceptionnel), est composée de représentants des ministères : de l’Intérieur, des Outre-Mer, de l’Économie et des Finances et de l’Écologie, du Développement durable et de l’Énergie. Les demandes de reconnaissance de l’état de catastrophe naturelle font l’objet d’une parution au Journal Officiel. Lorsque l’avis de la commission est favorable et confirmé par un arrêté interministériel, cela ouvre droit à une indemnisation au titre des contrats d’assurance. Par la suite, CCR indemnise l’assureur selon les modalités prévues par le traité de réassurance.

Les périls actuellement couverts par le régime sont les inondations (crues lentes, crues éclair, remontées de nappe, ruissellement, submersion marine), les coulées de boue, les séismes, les mouvements de terrain (y compris dues à la sécheresse), les affaissements de terrain dus à des cavités souterraines et à des marnières (sauf mines), les raz-de-marée, les

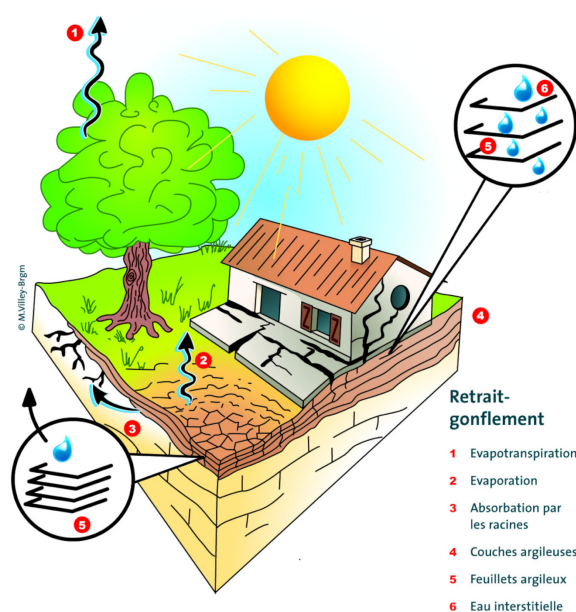


Figure 1.1: Le phénomène de retrait-gonflement des argiles provoque des fissures sur les maisons. La présence d'arbres à proximité du bâtiment est un facteur aggravant du fait de l'aspiration par les racines de l'eau du sol.

avalanches, les vents cycloniques de grande ampleur (supérieurs à 145 km/h en moyenne sur 10 mn ou 215 km/h en rafales). Cependant, cette liste n'est pas exhaustive et d'autres périls pourraient être indemnisés.

1.1.2 Péril sécheresse : définition en enjeux financiers

1.1.2.1 Phénomène de retrait-gonflement des argiles

Dans le cadre de cette thèse, le terme "sécheresse" désigne le phénomène qui entraîne l'apparition de désordres (principalement des fissures) dans les constructions individuelles (en grande majorité des maisons) suite aux mouvements différentiels des sols argileux et marneux. Ces mouvements sont provoqués par le gonflement de l'argile en présence d'humidité suivi de sa rétractation en période de sécheresse. Ce phénomène est aussi appelé "sécheresse RGA" ou tout simplement "RGA", pour retrait-gonflement des argiles. Il ne concerne pas les sécheresses agricoles qui ne sont pas prises en compte par le régime d'indemnisation des catastrophes naturelles français.

1.1.2.2 Facteurs intervenant dans le retrait-gonflement des argiles

La présence de certains facteurs, dits de prédisposition, permet ou favorise le phénomène de retrait-gonflement des argiles sans pour autant le déclencher. Le phénomène étudié

est provoqué lorsque la présence de facteurs de déclenchement s'ajoute à celle de facteurs de prédisposition.

Les facteurs de prédisposition : De par la nature du phénomène, la présence d'argile constitue le premier facteur de prédisposition. La pente au sol concourt également au retrait-gonflement des argiles puisqu'elle peut favoriser le ruissellement et le drainage. Également, les maisons construites sur un sol en pente sont plus souvent sujettes à une dissymétrie des fondations ce qui les rend plus vulnérables aux variations en teneur en eau du sol. La présence de nappes phréatiques de faible profondeur peut être à l'origine de variations de la teneur en haut des sols. Il en est de même pour la présence de végétation à proximité des maisons du fait du phénomène de succion correspondant à l'aspiration par les racines de l'eau du sol. Enfin, les défauts de construction représentent également un facteur de prédisposition, notamment lorsqu'ils concernent les fondations des maisons.

Les facteurs de déclenchement : Le facteur de déclenchement principal, d'ordre météorologique, correspond à l'alternance de périodes de sécheresse et d'humidité. Il existe un second facteur de déclenchement, d'ordre anthropique, correspondant à la réalisation de travaux de drainage et plus globalement d'aménagement. Ces réalisations modifient la répartition des écoulements souterrains et superficiels des eaux et peuvent provoquer des mouvements différentiels des sols.

1.1.2.3 Événement sécheresse

Un événement sécheresse correspond au phénomène de retrait-gonflement des argiles ayant eu lieu au cours d'une année civile. Dans le cadre de cette étude, le coût d'un événement sécheresse correspond au montant des dommages causés au cours d'une année civile par la sécheresse RGA à des bâtiments situés dans une communes bénéficiant de l'état de catastrophe naturelle au titre de ce péril et assurés par un assureur appartenant au portefeuille de CCR. La Caisse Centrale de Réassurance disposant d'une part de marché de 95% sur le marché français de la réassurance des catastrophes naturelles, le montant d'un événement sécheresse tel que défini précédemment présente une forte proximité avec le coût d'un événement sécheresse pour le marché français dans son intégralité.

1.1.2.4 Enjeux financiers

La sécheresse représente l'un des périls les plus coûteux, mais aussi l'un des moins connus du régime d'indemnisation des catastrophes naturelles. Pris en charge par le régime depuis 1989, ce péril représente 42% de la sinistralité globale hors automobile sur la période 1982-2022. Cela le place en deuxième position derrière les inondations en termes de dommages catastrophes naturelles sur cette période. En réalisant cette même analyse sur les 10 dernières années, la sécheresse apparaît être le péril le plus coûteux pour le régime d'indemnisation des catastrophes naturelles français avec 52% de la sinistralité globale [CCR, 2023]. La Figure 1.2 représente la sinistralité sécheresse en France de 1989

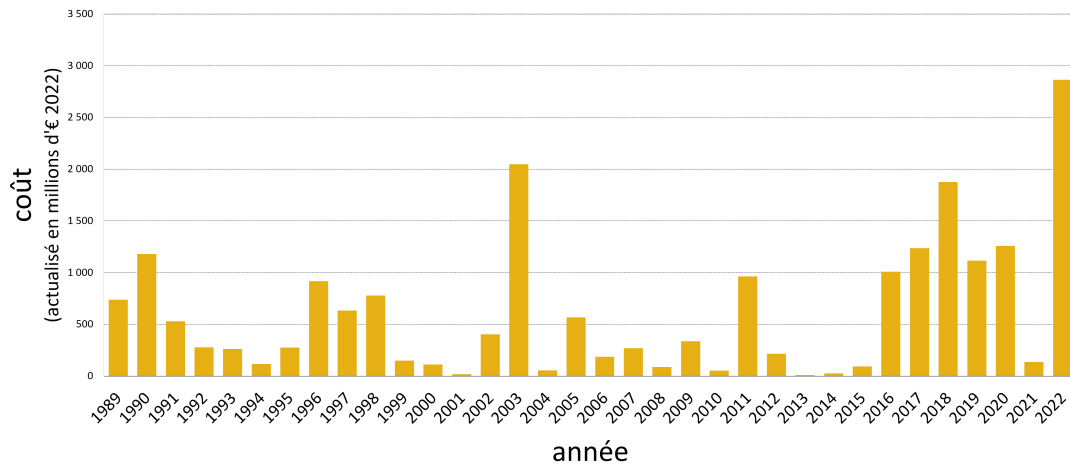


Figure 1.2: La sinistralité sécheresse en France de 1989 à 2022 (source CCR). Les montants affichés sont en millions d'euros et actualisés en euros 2022. Ces chiffres correspondent à l'ensemble du marché français et comprennent donc une part de sinistralité non couverte par CCR.

à 2022. Ainsi, les événements de sécheresse RGA peuvent provoquer une sinistralité importante, comme en 2022 (environ 3 Mds€) ou bien 2003 (2 Mds€ actualisés en euros 2022).

Par ailleurs, on constate une succession d'événements de grande ampleur entre 2016 et 2022. Sur cette période, la sinistralité moyenne s'élève à 1,35 Mds€ contre 610 M€ sur l'ensemble de la période 1989-2022. Malheureusement, les impacts financiers pourraient s'avérer encore supérieurs. En effet, d'après le Ministère de la Transition Écologique, 48% du territoire est en zone d'exposition moyenne ou forte au phénomène de retrait-gonflement des argiles [MTE, 2021].

La Figure 1.3 présente la carte d'exposition du territoire au phénomène de retrait-gonflement élaborée par le Bureau de Recherches Géologiques et Minières (BRGM). Elle résulte du croisement de données liées à la présence d'argile et de données liées à la sinistralité effectivement observée et a permis au Ministère de la Transition Écologique de quantifier le degré de vulnérabilité du parc des maisons individuelles français. Celle-ci apparaît particulièrement importante : 10 430 299 maisons individuelles sont situées en zone d'exposition moyenne ou forte, ce qui représente 54% de ces dernières [MTE, 2021]. Cette exposition du territoire français au RGA constitue un socle favorable aux effets du changement climatique sur la sinistralité sécheresse en France. Dans son étude de 2018, CCR anticipe une hausse de 23% à horizon 2050 de la sinistralité liée à ce péril [CCR, 2018].

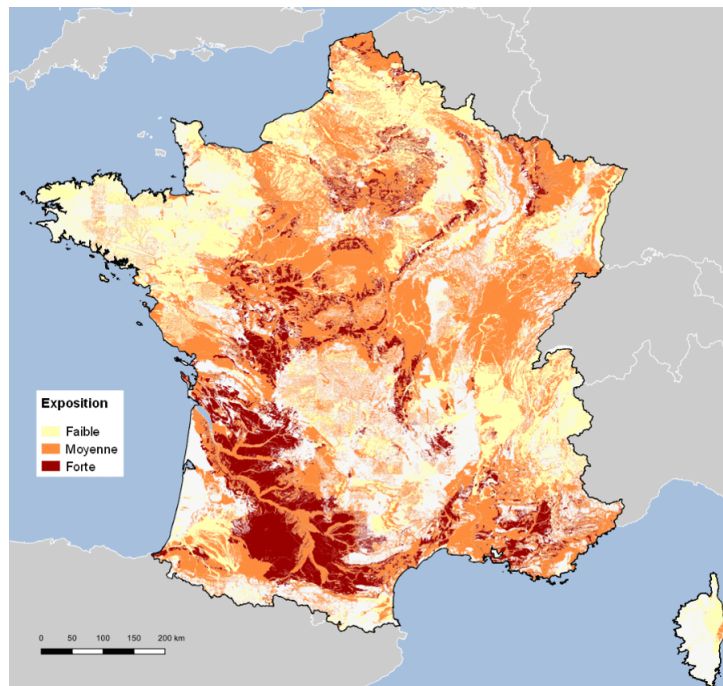


Figure 1.3: Cartographie de l'exposition du territoire au phénomène de retrait-gonflement : 48% du territoire est en zone d'exposition moyenne ou forte (source BRGM). Quatre catégories sont présentées et caractérisent l'exposition au RGA : à priori nulle en blanc, faible en jaune, moyenne en orange et forte en rouge.

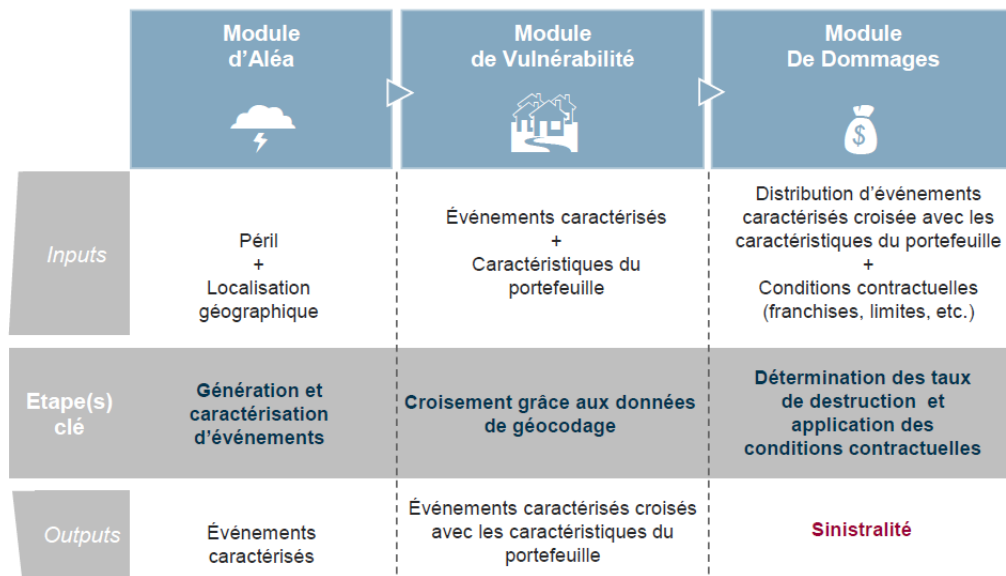


Figure 1.4: Les 3 modules d'un modèle catastrophe : aléa, vulnérabilité et dommages

1.1.3 De la nécessité d'anticiper les dommages

1.1.3.1 L'architecture d'un modèle catastrophes

Dans le but de mieux appréhender les risques liés aux catastrophes naturelles, CCR a développé (pour les inondations, la sécheresse, la submersion marine et les séismes) et adapté (pour les vents cyclones) des modèles catastrophe. Les modèles catastrophe sont tous articulés autour de trois modules :

- le module d'aléa génère un catalogue d'événements survenus ou fictifs, caractérisés par une probabilité de survenance et des données d'aléa relatives à l'intensité du phénomène physique étudié,
- le module de vulnérabilité recense l'ensemble des risques du portefeuille d'assurance étudié, avec leur localisation, leurs caractéristiques et leur valeur assurée et les croise avec les données d'aléa,
- le module de dommages, qui à partir du croisement entre les données d'aléa et de vulnérabilité permet d'estimer le montant des dommages et d'appliquer les conditions contractuelles.

La Figure 1.4 présente la décomposition d'un modèle catastrophe selon ces trois modules : aléa, vulnérabilité et dommages.

1.1.3.2 Les utilisations des modèles catastrophe

Les modèles catastrophe sont utilisés par CCR afin de quantifier sa propre exposition, celle de ses cédantes³ et celle de l'État français. Plus précisément, ces modèles permettent de tarifier les couvertures de réassurance proposées par CCR à ses cédantes. Ils sont également utiles pour le provisionnement, c'est-à-dire le calcul de la part de la prime en provenance des cédantes qu'il convient de mettre de côté afin d'être en mesure de faire face aux paiements futurs. Ici, il s'agit d'estimer le coût d'un événement sécheresse donné sur la base des observations relatives aux événements sécheresse l'ayant précédé. Par ailleurs, les modèles catastrophe sont mis à disposition pour le chiffrage des impacts des réformes du régime des catastrophes naturelles français proposées par l'État. Enfin, ces modèles permettent de quantifier l'impact du changement climatique sur le coût des catastrophes naturelles.

1.1.3.3 Modèle catastrophe déterministe et probabiliste

Il existe deux types de modèles catastrophe : les modèles catastrophe déterministes et les modèles catastrophe probabilistes. Dans leur version déterministe, les modèles catastrophe permettent d'estimer l'impact d'un événement dont les caractéristiques sont données. Dans leur version probabiliste, un catalogue d'événements probabilisés est créé et présenté au modèle catastrophes déterministe qui prédira les conséquences financières liées à chacun de ces événements.

Ces deux versions des modèles catastrophe répondent à des besoins différents. La version déterministe permet par exemple de chiffrer le coût d'une inondation récemment survenue. La version probabiliste permet notamment de tarifier une couverture de réassurance catastrophes naturelles : en sollicitant un catalogue d'aléa diversifié et des données de vulnérabilité liées au portefeuille d'assurés de la cédante, on obtient une distribution de la sinistralité de la cédante.

1.1.4 Objectifs de la thèse

L'objectif de cette thèse est de doter CCR d'un nouvel outil lui permettant d'anticiper de façon plus précise les dommages provoqués par la survenance d'un épisode de sécheresse RGA. Nous nous intéressons plus particulièrement aux dommages indemnisés dans le cadre du régime d'indemnisation des catastrophes naturelles français. Ainsi, l'estimation du coût d'un événement sécheresse nécessite en premier lieu de détecter les communes qui formuleront une demande de reconnaissance de l'état de catastrophe naturelle. Par la suite, il convient d'identifier celles d'entre elles qui sont éligibles aux critères de reconnaissance, et qui seront donc effectivement reconnues en état de catastrophes naturelles. La dernière étape consiste à estimer les dommages associés. L'estimation du coût de l'événement sécheresse correspond à la somme des estimations réalisées pour chaque commune.

3. Le terme cédante désigne un assureur client d'un réassureur

Pour un événement sécheresse donné, CCR est en mesure d'évaluer les critères de reconnaissance. Ainsi dans notre étude, nous nous plaçons dans la situation où les communes éligibles aux critères de reconnaissance sont connues.

Les dommages et les probabilités que chaque commune formule une demande de reconnaissance seront estimés à l'aide de Super Learners. Présenté dans la section suivante, le Super Learner constitue ainsi un fil conducteur de l'ensemble des travaux de cette thèse.

1.2 Super Learner

1.2.1 L'agrégation de modèles et le Super Learner

L'agrégation de modèles (ensemble learning en anglais) est une méthode d'apprentissage statistique reposant sur une collection d'algorithmes fondamentaux et visant à combiner les prédictions de chacun d'eux. Depuis la première approche proposée par Wolpert en 1992 [Wolpert, 1992a], de nombreux algorithmes ayant pour but d'agréger les prédictions d'algorithmes fondamentaux ont été développés, pouvant pour la plupart d'entre eux être regroupés en 3 catégories :

- Le bagging, qui repose sur l'apprentissage simultané de chaque algorithme fondamental sur des sous-échantillons différents. Les prédictions produites par les algorithmes fondamentaux sont ensuite agrégées en réalisant une moyenne ou éventuellement sur le principe du vote majoritaire dans le cas d'une classification. La forêt aléatoire [Breiman, 2001] est un exemple d'algorithme reposant sur le principe du bagging.
- Le boosting, qui repose sur l'apprentissage itératif des algorithmes fondamentaux, chacun d'entre eux visant à corriger l'erreur du précédent. Les prédictions produites par les algorithmes fondamentaux sont ensuite agrégées en réalisant une moyenne. Le gradient boosting [Friedman, 2001] est un exemple d'algorithme reposant sur le principe du boosting.
- Le stacking, qui repose sur une collection non contrainte d'algorithmes fondamentaux. Les prédictions de ces derniers sont agrégées en fonction des performances observées à la suite d'une validation croisée. Le Super Learner [van der Laan et al., 2007] et les algorithmes de type Robust online aggregation [Cesa-Bianchi and Lugosi, 2006, Littlestone and Warmuth, 1994] sont des exemples de modèles de stacking.

La Table 1.1 présente une vision synthétique des 3 principaux principes d'agrégation. Dans la suite, nous présentons plus en détail le Super Learner. Cet algorithme d'agrégation appartenant à la famille du stacking est à la base de l'algorithme développé dans le cadre de cette étude pour l'estimation des dommages liés à une sécheresse RGA en France.

Principe d'agrégation	Stratégie d'apprentissage	Stratégie d'agrégation	Exemple d'algorithme
Bagging	Apprentissage simultané sur des sous-échantillons différents	Moyenne, vote majoritaire	Forêt aléatoire [Breiman, 2001]
Boosting	Apprentissage itératif par correction successive des erreurs	Moyenne	Gradient boosting [Friedman, 2001]
Stacking	Souvent sans contrainte	Basée sur les résultats d'une validation croisée	Super Learner [van der Laan et al., 2007], Robust online aggregation [Cesa-Bianchi and Lugosi, 2006, Littlestone and Warmuth, 1994]

Table 1.1: Vue synthétique des 3 grands principes d'agrégation de modèles

1.2.2 Le Super Learner discret

Disposant d'un échantillon indépendamment et identiquement distribué (i.i.d.) $O_i = (Y_i, X_i) \in \mathcal{X} \times \mathbb{R}$ ($i = 1, \dots, n$) de loi \mathbb{P}_0 , nous nous intéressons à l'estimation d'un paramètre d'intérêt θ^* de la loi \mathbb{P}_0 . Ce paramètre est défini comme le minimiseur du risque associé à une fonction de perte ℓ sur un espace de paramètres Θ :

$$\theta^* = \arg \min_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_0}[\ell(\theta)(O)].$$

À titre d'exemple, si le paramètre d'intérêt θ^* correspond à l'espérance conditionnelle $\mathbb{E}_{\mathbb{P}_0}[Y|X]$ et Θ est un sous-ensemble de l'ensemble des fonctions de $\mathcal{X} \rightarrow \mathbb{R}$, alors la fonction de perte associée est l'erreur quadratique $\ell(\theta) : O \mapsto (Y - \theta(X))^2$.

Un algorithme $\hat{\theta}$ pour l'estimation de θ^* est formalisé comme une fonction qui associe à un jeu de données un estimateur de θ^* , que l'on nomme également prédicteur. En particulier, en désignant par \mathbb{P}_n la distribution empirique des observations $O_i, i = 1, \dots, n$, $\hat{\theta}(\mathbb{P}_n)$ estime θ^* . Le Super Learner discret sollicite une collection de J algorithmes $\hat{\theta}_1, \dots, \hat{\theta}_J$ et sélectionne l'algorithme $\hat{\theta}_{\hat{j}}$ présentant les meilleures performances telles qu'évaluées dans le cadre d'un schéma de validation, par exemple une validation croisée à V volets :

$$\hat{j} = \arg \min_{j \in J} \frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_{n,V(v)}}[\ell(\hat{\theta}_j(\mathbb{P}_{n,T(v)}))(O)],$$

où $V(v)$ et $T(v)$ correspondent aux indices des observations appartenant respectivement au $v^{\text{ème}}$ volet (données de validation) et à l'ensemble des données privé de celles-ci (données d'apprentissage). Ici, pour tout $v \in 1, \dots, V$, $\mathbb{P}_{n,V(v)}$ et $\mathbb{P}_{n,T(v)}$ correspondent aux distributions empiriques des sous-échantillons $\{O_i : i \in V(v)\}$ et $\{O_i : i \in T(v)\}$.

L'algorithme oracle est l'algorithme bénéficiant des meilleures performances au terme de la validation croisée sous la loi \mathbb{P}_0 :

$$\tilde{j} = \arg \min_{j \in J} \frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} [\ell(\hat{\theta}_j(\mathbb{P}_{n,T(v)}))(O)].$$

Ainsi, l'algorithme oracle est en pratique inconnu.

Une justification théorique de la procédure de sélection du Super Learner discret est présentée dans [van der Laan et al., 2007, Theorem 2]. La clef de l'analyse réside dans la proximité entre $\mathbb{E}_{\mathbb{P}_{n,V}}[\ell(\hat{\theta}_j(\mathbb{P}_{n,T(v)}))(O)]$ et $\mathbb{E}_{\mathbb{P}_0}[\ell(\hat{\theta}_j(\mathbb{P}_{n,T(v)}))(O)]$:

Théorème 1 (Inégalité oracle pour le Super Learner discret). *Soient, pour tout $\theta \in \Theta$,*

$$\Delta^* \ell(\theta) := \ell(\theta) - \ell(\theta^*) \quad \text{et} \quad d_0(\theta, \theta^*) := \mathbb{E}_{\mathbb{P}_0}[\Delta^* \ell(\theta)(O)],$$

les différences des pertes et des risques de θ et θ^ . Considérons les deux hypothèses suivantes.*

Hypothèse 1. *Il existe un réel $M_1 < \infty$ tel que*

$$\sup_{\theta \in \Theta} \|\Delta^* \ell(\theta)\|_\infty \leq M_1.$$

Hypothèse 2. *Il existe un réel $M_2 < \infty$ tel que*

$$\sup_{\theta \in \Theta} \frac{\text{var}_{\mathbb{P}_0}[\Delta^* \ell(\theta)(O)]}{d_0(\theta, \theta^*)} \leq M_2.$$

Supposons que pour tout $j \in \llbracket 1, J \rrbracket$, $\mathbb{P}(\hat{\theta}_j(\mathbb{P}_n) \in \Theta) = 1$. Alors sous les Hypothèses 1 et 2, pour tout $\lambda > 0$:

$$\frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_{\tilde{j}}(\mathbb{P}_{n,T(v)}), \theta^*) \leq (1 + 2\lambda) \frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_{\tilde{j}}(\mathbb{P}_{n,T(v)}), \theta^*) + C(\lambda) \frac{V(1 + \ln(J))}{n}, \quad (1.1)$$

où $C(\lambda) := 16(M_1 + \frac{3(1+\lambda)}{2\lambda} M_2)(1 + \lambda)$. Par conséquent, sous ces mêmes hypothèses, il existe une constante $C' > 0$ telle que

$$\frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_{\tilde{j}}(\mathbb{P}_{n,T(v)}), \theta^*) \leq \frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_{\tilde{j}}(\mathbb{P}_{n,T(v)}), \theta^*) + C' \sqrt{V \frac{1 + \ln(J)}{n}} \quad (1.2)$$

dès lors que $V(1 + \ln(J))/n$ est suffisamment petit.

La preuve du Théorème 1, qui repose notamment sur une inégalité de concentration de type Bernstein [Boucheron et al., 2013], est disponible dans [van der Laan et al., 2006].

L'inégalité (1.2) déduite de (1.1) n'apparaît pas dans [van der Laan et al., 2006] mais sa preuve est simple. En effet, (1.1) s'écrit

$$\frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_j(\mathbb{P}_{n,T(v)}), \theta^*) \leq \frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_j(\mathbb{P}_{n,T(v)}), \theta^*) + \text{reste}(\lambda) \quad (1.3)$$

avec

$$\begin{aligned} \text{reste}(\lambda) &= 2\lambda \frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_j(\mathbb{P}_{n,T(v)}), \theta^*) + C(\lambda) V \frac{1 + \ln(J)}{n} \\ &\leq 2\lambda M_1 + 16 \left(M_1 + \frac{3(1 + \lambda)}{2\lambda} M_2 \right) (1 + \lambda) V \frac{1 + \ln(J)}{n}. \end{aligned}$$

Soit $a, b, c, d = V \frac{1 + \ln(J)}{n}$ quatre constantes positives telles que $\text{reste}(\lambda) = a\lambda + (b + c \frac{1 + \lambda}{\lambda})(1 + \lambda)d$ pour tout $\lambda > 0$. Supposons que $d \leq 1$ (soit, heuristiquement, que n est grand) et choisissons $\lambda = \sqrt{d}$. Alors

$$\begin{aligned} \text{reste}(\lambda) &= a\sqrt{d} + \left(b + c \frac{1 + \sqrt{d}}{\sqrt{d}} \right) (1 + \sqrt{d})d \\ &\leq a\sqrt{d} + 2 \left(b + c \frac{2}{\sqrt{d}} \right) d \\ &= a\sqrt{d} + 2bd + 4c\sqrt{d} \\ &\leq (a + 2b + 4c)\sqrt{d}. \end{aligned}$$

En injectant cette inégalité dans (1.3), on obtient bien (1.2) pour $C' = a + 2b + 4c$. Ce théorème indique que le Super Learner discret est asymptotiquement aussi performant au sens de d_0 que le sélecteur oracle, à une constante près, dès lors que $V(1 + \ln(J))/n$ est suffisamment petit et que la performance du sélecteur oracle domine ce quotient.

1.2.3 Le Super Learner continu

Le Super Learner décrit dans la section précédente sélectionne le meilleur algorithme θ^* au sein d'une collection de J algorithmes $\hat{\theta}_1, \dots, \hat{\theta}_J$. Le Super Learner discret se base sur la validation croisée afin d'évaluer les performances des algorithmes.

Plutôt que de sélectionner l'un d'entre eux, le Super Learner continu introduit dans [van der Laan et al., 2007] permet de combiner les prédictions des prédicteurs. Si l'algorithme agrégateur noté $\hat{\theta}^{SL}$ est au choix de l'utilisateur (il peut s'agir par exemple d'un algorithme de machine learning), des garanties théoriques sont proposées dans le cas où l'agrégateur est paramétrique et indexé par un paramètre de dimension finie $\alpha \in \mathcal{A}$. Plus spécifiquement, nous nous intéressons ici au cas où l'agrégateur réalise une combinaison convexe des prédictions des prédicteurs. Pour $\mathcal{A} = \{\alpha_1, \dots, \alpha_J \in [0, 1] : \sum_{i=1}^J \alpha_i = 1\}$, le Super Learner continu est de la forme :

$$\hat{\theta}_\alpha^{SL}(\mathbb{P}_n) = \sum_{i=1}^J \alpha_i \hat{\theta}_i(\mathbb{P}_n).$$

Dans la pratique, l'ensemble \mathcal{A} est discrétisé en un ensemble fini \mathcal{A}_n de valeurs de \mathcal{A} . L'implémentation de ce Super Learner continu est la suivante :

1. Entraîner les J algorithmes sur l'ensemble des données. À ce stade nous disposons des prédicteurs $\hat{\theta}_1(\mathbb{P}_n), \dots, \hat{\theta}_J(\mathbb{P}_n)$.
2. Répartir les données en V sous-échantillons disjoints de même taille n_v .
3. Pour chaque v , entraîner l'ensemble des J algorithmes sur le sous-échantillon indexé par $T(v)$ correspondant. À ce stade nous disposons des $J \times V$ prédicteurs $\hat{\theta}_j(\mathbb{P}_{n,T(v)})$, $j = 1, \dots, J$, $v = 1, \dots, V$.
4. Réaliser les prédictions $\hat{\theta}_j(\mathbb{P}_{n,T(v)})(X_i)$, $j = 1, \dots, J$, $v = 1, \dots, V$, $i \in V(v)$.
5. Déterminer les poids $\hat{\alpha} \in \mathcal{A}_n$ tels que

$$\hat{\alpha} = \arg \min_{\alpha \in \mathcal{A}_n} \frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_{n,V(v)}}[\ell(\hat{\theta}_\alpha^{SL}(\mathbb{P}_{n,T(v)}))(O)].$$

Si la fonction de perte ℓ correspond à l'erreur quadratique, alors

$$\hat{\alpha} = \arg \min_{\alpha \in \mathcal{A}_n} \frac{1}{V} \sum_{v=1}^V \frac{1}{n_v} \sum_{i \in V(v)} (Y_i - \hat{\theta}_\alpha^{SL}(\mathbb{P}_{n,T(v)})(X_i))^2.$$

6. Finalement, $\hat{\theta}^{SL}(\mathbb{P}_n) = \hat{\theta}_\alpha^{SL}(\mathbb{P}_n)$.

Dans le cadre du Super Learner continu, le sélecteur oracle sollicite le même ensemble de poids discrétisé \mathcal{A}_n mais retient les poids permettant de minimiser le risque sous la loi \mathbb{P}_0 :

$$\tilde{\alpha} = \arg \min_{\alpha \in \mathcal{A}_n} \frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0}[\ell(\hat{\theta}_\alpha^{SL}(\mathbb{P}_{n,T(v)}))(O)].$$

Au même titre que sous sa forme discrète, le Super Learner continu bénéficie de garanties théoriques d'apprentissage. Le théorème suivant est issu de [van der Laan et al., 2007] :

Théorème 2 (Inégalité oracle pour le Super Learner continu). *Supposons qu'il existe un ensemble borné $\mathcal{Y} \in \mathbb{R}$ et un ensemble Euclidien borné \mathcal{X} tels que $\mathbb{P}_0((Y, X) \in \mathcal{Y} \times \mathcal{X}) = 1$ et $\mathbb{P}_0(\hat{\theta}_j(\mathbb{P}_n)(O) \in \mathcal{Y}) = 1$. Pour chaque $\delta > 0$, il existe une constante $C(\delta) < \infty$ telle que*

$$\frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_\alpha^{SL}(\mathbb{P}_{n,T(v)}), \theta^*) \leq (1 + \delta) \frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_\alpha^{SL}(\mathbb{P}_{n,T(v)}), \theta^*) + C(\delta) \frac{V \ln(n)}{n}.$$

Sous la condition supplémentaire que

$$\frac{\ln(n)}{n \frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_\alpha^{SL}(\mathbb{P}_{n,T(v)}), \theta^*)} \rightarrow 0 \text{ lorsque } n \rightarrow \infty,$$

nous obtenons

$$\frac{\frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_\alpha^{SL}(\mathbb{P}_{n,T(v)}), \theta^*)}{\frac{1}{V} \sum_{v=1}^V \mathbb{E}_{\mathbb{P}_0} d_0(\hat{\theta}_\alpha^{SL}(\mathbb{P}_{n,T(v)}), \theta^*)} \rightarrow 1 \text{ lorsque } n \rightarrow \infty.$$

Ce théorème portant sur l'analyse du Super Learner continu s'interprète de façon analogue au Théorème 1 pour le Super Learner discret. Toutefois, si l'ensemble des poids \mathcal{A}_n comprend les J éléments de \mathcal{A} affectant un poids de 1 au j^e algorithme de la collection et 0 aux autres, alors le Théorème 2 indique que le Super Learner continu présentera des performances asymptotiquement meilleures ou équivalentes à celles de n'importe quel algorithme de la collection.

1.2.4 Le Super Learner séquentiel

Qu'il soit discret ou continu, le Super Learner a été présenté dans les sections précédentes sous sa forme originale. Depuis, plusieurs extensions ont été proposées et étudiées (voir à titre d'exemple [Sapp et al., 2014] ou encore [Wu and Benkeser, 2022]). Le Super Learner en ligne introduit dans [Benkeser et al., 2018] représente une extension d'intérêt pour l'étude de la sécheresse RGA puisqu'il tient compte de l'arrivée successive des observations, de la dépendance entre celles-ci, et de la nécessité de réaliser à l'instant $t \in \mathbb{N}^*$ des estimations relatives à l'instant $(t+1)$ en disposant des observations relatives aux pas de temps t et antérieurs. Cette configuration correspond notamment au contexte du provisionnement évoqué en Section 1.1.3.2.

Dans le cadre de notre étude, les estimateurs sollicités ne sont pas en ligne dans la mesure où le volume des données ne le nécessite pas. Nous parlerons donc de Super Learner séquentiel dans la suite de ce document.

Au pas de temps $t \in \mathbb{N}^*$ nous disposons des observations O_1, \dots, O_t . Pour chaque $t \in \mathbb{N}^*$, $F_{t-1} = \sigma(O_{i,\tau}, 1 \leq i \leq n, 1 \leq \tau < t)$ est la filtration engendrée par les observations du passé. Enfin, au pas de temps $t \geq 1$ chaque algorithme $\hat{\theta}_j$, $j = 1, \dots, J$, produit un estimateur $\theta_{j,t}$ de θ^* .

Dans sa forme discrète, le Super Learner séquentiel sélectionne l'algorithme $\hat{\theta}_{\hat{j}_t}$ présentant les meilleures performances à la suite d'une validation croisée séquentielle :

$$\hat{j}_t = \arg \min_{j \in J} \frac{1}{t} \sum_{\tau=1}^t \ell(\theta_{j,\tau-1})(O_\tau).$$

Le sélecteur oracle retient le meilleur algorithme $\hat{\theta}_{\tilde{j}_t}$ en évaluant séquentiellement les performances des prédicteurs sous la loi \mathbb{P}_0 conditionnellement aux informations passées résumées par F_t :

$$\tilde{j}_t = \arg \min_{j \in J} \frac{1}{t} \sum_{\tau=1}^t \mathbb{E}_{\mathbb{P}_0}[\ell(\theta_{j,\tau-1})(O_\tau) | F_{\tau-1}].$$

De même que dans le cas i.i.d. il existe une version continue du Super Learner séquentiel. Par ailleurs des garanties d'apprentissage sont proposées pour cette extension du Super Learner, à nouveau sous la forme d'inégalité oracle [voir Benkeser et al., 2018, section 4.2].

1.3 Apports de la thèse

1.3.1 Forecasting the cost of drought events in France by Super Learning from a short time series of many slightly dependent data

Le deuxième chapitre de cette thèse est consacré à l'anticipation des dommages liés à un événement sécheresse en France. Un algorithme original a été développé et étudié. Celui-ci se base sur le Super Learner, un algorithme d'agrégation de modèles présenté en Sections 1.2.2 et 1.2.3. Fondé sur une collection d'algorithmes fondamentaux fournie par l'utilisateur, le Super Learner a recours à une validation croisée pour évaluer chacun d'eux. À la suite de cette validation le Super Learner, dans sa version discrète, retient le meilleur algorithme ou bien, dans sa forme continue, constitue ses prédictions comme une combinaison linéaire convexe des prédictions de chaque algorithme. L'algorithme développé dans le cadre de cette étude se distingue du Super Learner originel dans la mesure où, plutôt que d'agrèger les prédictions des algorithmes fondamentaux, il agrège les prédictions en provenance d'une nouvelle collection composée de meta-algorithmes. Ici, chacun de ces meta-algorithmes agrège lui-même les prédictions des algorithmes fondamentaux. Comme dans le cadre du Super Learner séquentiel présenté en Section 1.2.4, une validation séquentielle est déployée. Cependant, dans le cadre de cette étude, nous observons à chaque pas de temps $t \in \mathbb{N}^*$ le vecteur \bar{O}_t composé des observations spatialement dépendantes $(O_{\alpha,t})_{\alpha \in \mathcal{A}}$, où \mathcal{A} correspond à l'ensemble des communes de France métropolitaine. Dans le cadre de l'analyse du Super Learner original développé, nous établissons qu'il est possible d'apprendre un paramètre θ^* de la loi des observations malgré le faible nombre d'observations de notre série temporelle (une trentaine d'événements sécheresse sont observés). Le déploiement de l'algorithme proposé pour l'estimation des dommages liés à une sécheresse a nécessité la constitution d'un jeu de données décrivant à la fois les communes de France métropolitaines et les différents événements de sécheresse observés. Le jeu de données exploité a été construit par le croisement de données en provenance de plusieurs sources telles que les cédantes de CCR, l'Institut National de la Statistique et des Études Économiques (INSEE), le Bureau de Recherches Géologiques et Minières (BRGM), l'Institut Géographique National (IGN) ou encore Météo-France. Par la suite, deux collections diversifiées d'algorithmes ont été constituées pour l'implémentation de la procédure proposée, la première constituée d'algorithmes fondamentaux et la seconde constituée de meta-algorithmes. Enfin, une étude portant sur l'importance des variables a été menée dans le but de faciliter l'acceptabilité de l'algorithme développé dans ce chapitre.

Ce chapitre est le fruit d'une collaboration avec Aurélien Bibaut (qui était alors affilié à la Division of Biostatistics de UC Berkeley) et Antoine Chambaz (le directeur de cette thèse). Ce chapitre a fait l'objet de deux prépublications [Ecoto et al., 2021b, Ecoto and Chambaz, 2022b]. Combinées en un unique manuscrit, celles-ci sont soumises à une revue internationale.

1.3.2 L'anticipation des communes demanderesses de la reconnaissance de l'état de catastrophe naturelle par Super Learning

Dans le troisième chapitre de cette étude, une instance de l'algorithme développé et étudié précédemment est déployée afin d'estimer les probabilités de demande de reconnaissance de l'état de catastrophe naturelle de chaque commune au titre d'un événement sécheresse donné. Les performances seront par la suite améliorées par la prise en compte d'une nouvelle source de données transmise par la commission interministérielle et ouvrant la voie à une approche originale pour l'estimation des probabilités de demande de reconnaissance. Ces nouvelles données n'étant disponibles de façon fiable qu'à partir de 2019, l'historique considéré est alors considérablement réduit. S'inspirant du deep learning, le transfer learning permettra de bénéficier à la fois de l'information contenue dans l'historique précédent cette date, et à la fois de l'information enrichie des nouvelles données. Les prédictions se verront par ce biais à nouveau améliorées.

1.3.3 Making sparse predictions, and forecasting the requests of the government declaration of natural disaster for a drought event in France

Le quatrième chapitre de cette étude présente un nouvel algorithme développé pour l'estimation des demandes de reconnaissance de l'état de catastrophe naturelle. Cet algorithme repose sur les prédictions d'un Super Learner pour la quantification du nombre de demandes de reconnaissance à l'échelle nationale. Les demandes sont ensuite réparties par communes en ayant recours au transport optimal. Les prédictions finales correspondent à la moyenne géométrique entre les prédictions réparties par transport optimal et les prédictions initiales du Super Learner. Une étude de simulations illustre les performances de l'algorithme proposé. Enfin, l'application aux données réelles fait ressortir des performances supérieures à celles du Super Learner initial.

Ce chapitre est le fruit d'une collaboration avec Thi Thanh Yen Nguyen (doctorante au laboratoire MAP5) et Antoine Chambaz (notre directeur de thèse). Ce chapitre fera prochainement l'objet d'une prépublication. Il sera alors soumis à une revue internationale.

1.3.4 Conclusion et perspectives

Le dernier chapitre de cette étude résume les conclusions des travaux réalisés et présente des pistes de recherches.

Chapter 2

Forecasting the cost of drought events in France by Super Learning from a short time series of many slightly dependent data

Ce chapitre est le fruit d'une collaboration avec Aurélien Bibaut (qui était alors affilié à la Division of Biostatistics de UC Berkeley) et Antoine Chambaz (mon directeur de thèse). Du fait de ma maîtrise du problème considéré, j'ai joué un rôle déterminant dans sa modélisation, dans la préparation des données, dans la conception de l'algorithme, dans sa programmation, dans son application et dans l'analyse des résultats.

Ce chapitre a fait l'objet de deux prépublications [Ecoto et al., 2021b, Ecoto and Chambaz, 2022b]. Combinées en un unique manuscrit, celles-ci sont soumises à une revue internationale.

2.1 Introduction

In this study we call *a drought event* the phenomenon of clay shrinking and swelling during a calendar year. We refer to [Charpentier et al., 2022, Sections 1 and 2] for an excellent introduction to drought events and their economic consequences. In a nutshell, the clay present in the soil alternatively shrinks and swells in dry and humid conditions. This creates instabilities and generates cracks in buildings. The cracks induce costs covered by private property insurance policies. Because 90% of the French natural disasters insurance market is reinsured by [Caisse Centrale de Réassurance](#) (henceforth abbreviated as CCR) [CCR, 2022b], a public-sector reinsurer providing cedents⁴ operating in France with coverage against natural catastrophes and uninsurable risks, the

4. A cedent is a party in an reinsurance contract that passes the financial obligation for certain potential losses to the reinsurer. In return for bearing a particular risk of loss, the cedent pays a reinsurance premium.

French state is eventually exposed.

France has been facing severe drought events over the past years. According to CCR [CCR, 2021], the average annual cost of drought events between 2016 and 2020 is 1.1 billion euros, a threefold increase relative to the 2002-2015 period (in the aforementioned reference and in the present study, unless stated otherwise, euros are constant euros). In this light, the recent cycle of extremely intense drought events raises two questions: will climate change perpetuate this pattern [Bradford, 2000, Iglesias et al., 2019] and, if so, what cost will the French state incur?

In a long-term perspective, CCR has studied the impact of climate change on the damages caused by natural disasters based on the Intergovernmental Panel on Climate Change (IPCC) scenarios RCP 4.5 and RCP 8.5 [CCR, 2015, 2018]. Resorting to ARPEGE simulations of the climate in 2050 provided by Météo-France, CCR simulated damages in France in 2050 and concluded that the annual cost of drought events in 2050 could increase, depending on the scenario, by 3% (under scenario RCP 4.5) or 23% (under scenario RCP 8.5). Unfortunately, the latter is more likely today than the former.

In a short- to middle-term perspective, forecasting the cost of drought events in France is an important task for CCR. Due to intricacies of the French legal framework [known as the natural disasters compensation scheme, see Charpentier et al., 2022, Section 2.1], the cost of a drought event on a given year is the overall aggregate cost across the cities that obtained the government declaration of natural disaster for a drought event that year. We stress that we used and will use from now on the word city as a translation of the French word *commune*, a level of administrative division in France. We use the word city irrespective of the *commune*'s size, which can vary widely from small hamlets to large cities. Going back to the forecasting task, it will be carried out several times every year because, as months goes by, more relevant information is accrued. At first, it is necessary to predict which cities will make a request for the government declaration of natural disaster for a drought event. Later on it is known that some cities did make the request but it is still necessary to make predictions for the others. Later still it is known exactly which cities made the request. Note that once a request is made, there is no uncertainty for CCR about whether or not the city will obtain the government declaration of natural disaster for a drought event. Therefore CCR currently addresses two sub-problems separately: sub-problem 1 consists in predicting which cities will make a request for the government declaration of natural disaster for a drought event; sub-problem 2 consists in predicting the cost of a drought event for those cities that obtained the government declaration of natural disaster for a drought event. In this study, we focus on sub-problem 2. On the contrary, [Chatelain and Loisel, 2021] addresses the two sub-problems. As for [Charpentier et al., 2022, Heranval et al., 2022], they predict which cities will experience claims (a proxy to sub-problem 1) and then the cost for these cities. We acknowledge that the problem we tackle is therefore more circumscribed than theirs. Quoting [Logar and van den Bergh, 2011, page 4, first paragraph], “[t]he existing literature on the costs of drought [events] is scarce, fragmented and heterogeneous and there is a need for comprehensive costs estimations to help designing effective policy responses.”

To the best of our knowledge, [Charpentier et al., 2022, Heranval et al., 2022, Chatelain and Loisel, 2021] are the only three references available about the prediction of the cost of drought events. The studies conducted by insurance companies are confidential. In [Charpentier et al., 2022], the authors use Generalized Linear Models (GLM) and tree-based machine learning algorithms (variants of the random forest algorithm). For a given drought event, for each city, the number of claims and the average cost are predicted, then a city-specific predicted cost is obtained by multiplying these two numbers. The overall cost is finally estimated by the sum of all the city-specific costs. In [Heranval et al., 2022], the authors use penalized GLM and machine learning algorithms (random forests and extreme gradient boosting) to predict which cities will experience claims. For a given drought event, for each city susceptible to experience claims, they then use a common linear regression model to map the number of houses to a city-specific cost. The overall cost is finally estimated by the sum of these city-specific costs. In [Chatelain and Loisel, 2021], the authors use GLM and the extreme gradient boosting algorithm to predict which cities will make a request for the government declaration of natural disaster for a drought event. For a given drought event, for each city susceptible to make a request, they then use several GLM to predict house-level costs (using geolocated data).

In the present study, we develop and apply a new methodology to forecast the cost of a drought event in France. In brief, we model the data set as a time series where each time- t -specific data-structure ($t \in \mathbb{N}^*$ represents a year) is made of dependent data $O_{\alpha,t}$ further indexed by $\alpha \in \mathcal{A}$ (α represents a city). Each $O_{\alpha,t}$ decomposes as $(C_{\alpha,t}, W_{\alpha,t}, Y_{\alpha,t})$ where $C_{\alpha,t}$ is a collection of (α, t) -specific covariates, $W_{\alpha,t} \in \{0, 1\}$ is an indicator of whether or not the city α obtained the government declaration of natural disaster for a drought event on year t ($W_{\alpha,t} = 1$ if it did), and $Y_{\alpha,t}$ is the (α, t) -specific cost of the drought event. Assuming that the conditional expectation θ^* of the cost $Y_{\alpha,t}$ given the covariates $C_{\alpha,t}$, $W_{\alpha,t} = 1$ and the past does not depend on (α, t) and the past (a stationarity assumption), we propose to build an estimator θ_{t-1} of θ^* using all data till time $(t-1)$. We then predict the overall cost at time t , $\sum_{\alpha \in \mathcal{A}} Y_{\alpha,t} \mathbf{1}\{W_{\alpha,t} = 1\}$, with $\sum_{\alpha \in \mathcal{A}} \theta_{t-1}(C_{\alpha,t}) \mathbf{1}\{W_{\alpha,t} = 1\}$.

The estimation of θ^* is made using a learning algorithm that builds upon a library of competing algorithms, and either selects the one that performs best or combines the algorithms into a single meta-algorithm that performs almost as well as all possible combinations thereof. This is known as stacking, or aggregating, or super learning in the literature [van der Laan et al., 2007], [Polley et al., 2011, and references therein]. We call our learning algorithm the One-Step Ahead Sequential Super Learner (OSASSL). As in [Benkeser et al., 2018], the OSASSL respects the temporal structure of the data set.

Like in [Charpentier et al., 2022, Heranval et al., 2022], we exploit the Soil Wetness Index (SWI) as a drought indice [it is referred to as the Standardised Soil Water Index by Charpentier et al., 2022]. Moreover, like Charpentier et al. [2022], we also use sequential cross-validation to take into account the time dependence structure in our data set. In contrast to [Charpentier et al., 2022, Heranval et al., 2022], we rely on a richer description of the cities which we obtained by data enrichment (more details to follow).

We face several challenges. From the applied point of view, assembling the learning data set is difficult because the data come from many sources and take on various shapes. Moreover, some of the data are only partially available. From a theoretical point of view, the time series is observed only at a limited number of time steps and each t -specific observation is a complex data-structure with an intricate dependence pattern. Could the shortness of the time series be compensated by the large cardinality of \mathcal{A} , $|\mathcal{A}|$, provided that there is a large amount of independence among the dependent $O_{\alpha,t}$ ($\alpha \in \mathcal{A}$)? This question motivates an original theoretical analysis that uses dependency graph to model the amount of conditional independence within each t -specific data-structure and a concentration inequality by Janson [2004] in order to leverage a large ratio of $|\mathcal{A}|$ to the degree of the dependency graph in the face of a small number of t -specific data-structures.

In Section 2.2, we present the data that we collected and used in this study. In Section 2.3, we give a brief historical perspective of the concept of aggregation, describe the OSASSL and present a theoretical result about its performance. The result follows from an analysis that we fully expose in the Appendix. In Section 2.4, we present and comment on the numerical results that we obtain. We make comparisons with other aggregation strategies from the robust online aggregation literature [Cesa-Bianchi and Lugosi, 2006] and assess how the covariates used to predict the costs influence the predictions. In Section 2.5, we discuss directions for future work.

2.2 Data

We merge several data sets into a master data set. The merged data sets are either provided by CCR's cedents or are collected by us from other sources. They contribute different kinds of information.

Of note, in the rest of this study, France refers to *Metropolitan* or *Mainland* France. Drought events are not a threat in Overseas France (essentially because there is little clay in these parts of the country).

2.2.1 Data provided by CCR's cedents

Of note, 90% of the French natural disasters insurance market is reinsured by CCR [CCR, 2022b]. Contractually, its cedents must share their portfolios and claims data. Over the years, CCR has thus gathered a vast collection of accurate localizations and characteristics of insured goods and claims data. From 1990 to present, the collection covers roughly 22% to 97% of the overall cost of all the French claims (see Section 2.2.3.1 and Figure 2.1).

2.2.2 Data garnered from other sources

The data set, based so far on data provided by cedents only, is then enriched with data from four trusted public organizations that collect, share and analyze information about

the French economy and people (National Institute for Statistical and Economic Studies, Insee), geography (Geographic National Institute, IGN), geology (French Geological Survey, BRGM) and meteorology (Météo-France). The new features supplementing the description of the cities are seismic and climatic zones, clay shrinkage-swelling hazards, tree-coverage rate, area, population and years of construction. Lastly, we benefit from the Soil Wetness Index (SWI) as described in [Dirmeyer et al., 1999] and in [this document](#) made available by Météo-France.

2.2.3 City-level data processing

Some data are available at the house-level (namely, the cost of claim and insured sum), but most are not. In particular, the pivotal SWI variable is available at a 8×8 km² resolution, while the 90%-quantile of the French cities area is 30 km² and only 1.3% of the French cities have an area larger than 65 km² (data from 2014). Consequently, we choose to work at a city level and thus aggregate the features that have a higher resolution. Details follow.

2.2.3.1 On the city-level costs of drought events

The cost of the damages in a city caused by a drought event (what will be our response variable) is unknown. However, on the one hand the overall cost across France is estimated (in both current and constant euros) by actuarial studies conducted by CCR and, on the other hand, we know the costs of *those* claims filled in the claims data provided by the cedents which, unfortunately, only represent a fraction of all the claims.

Provisional city-specific costs are computed by aggregating by city the costs filled in the claims data provided by the cedents. Because these claims data are not exhaustive, the sum of all the provisional city-specific costs is smaller than the estimated overall cost. The (final) city-specific costs are proportional to the provisional city-specific costs in such a way that the sum of all the (final) city-specific costs equals the estimated overall cost in constant euros.

Figure 2.1 illustrates the gaps between the estimated overall costs across France and the sum of the provisional city-specific costs. The ratios of the latter to the former range from 22% to and 97% (the 144%-ratio corresponds to an exceptional year where the estimated overall cost is even smaller than the very small aggregated cost of the claims data).

2.2.3.2 On the city-level SWI

The clay present in the soil shrinks and swells in dry and humid conditions, creating instabilities and generating cracks in buildings. In order to quantify the soil humidity we use the SWI. Provided by Météo-France, the SWI data consist of time series of values (one every ten-day period, a *décade* in French) ranging between -3.33 (very dry soil) and 2.33 (very wet soil). Figure 2.2 presents five one-year chunks of SWI time series.

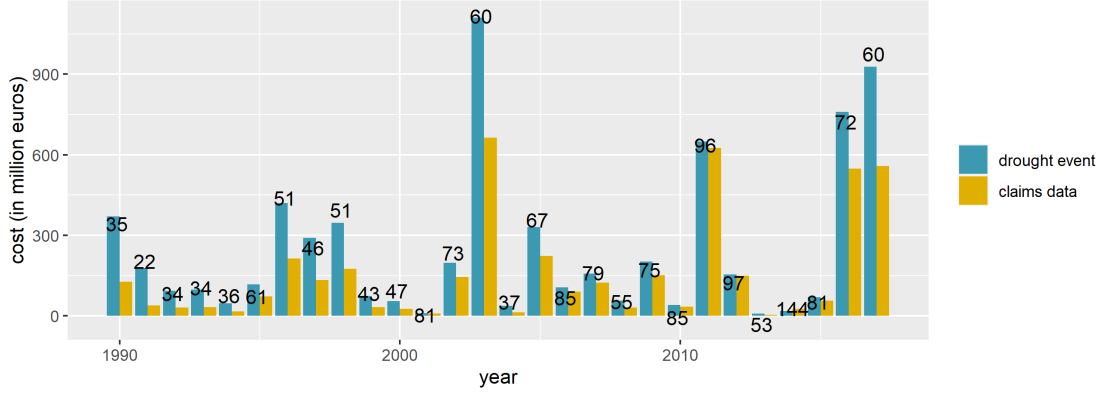


Figure 2.1: *Estimated overall costs of drought events across France (blue) and provisional city-specific costs obtained by aggregating the costs of those claims filled in the claims data provided by the cedents (yellow). The ratios of the latter to the former (numbers above the blue bars) range between 22% and 97% (the 144%-ratio corresponds to an exceptional year where the estimated overall cost is even smaller than the very small aggregated cost of the claims data). In this figure we use current euros. Source: CCR.*

For every year and every city, we derive a collection of 36 city-level SWIs, one for each ten-day period that make up a year. Each of these 36 SWIs is the convex average of the corresponding SWIs of the $8 \times 8 \text{ km}^2$ squares that overlap the city's area. The weights are proportional to the areas of the intersections.

We use the city-level SWIs to build a rich collection of SWI-related covariates.

Because the effects of a drought event can build up slowly [Logar and van den Bergh, 2011, page 10, second paragraph], for every year t and every city, we concatenate the 3×36 ten-day city-level SWIs of years t , $(t-1)$ and $(t-2)$. We also add the minima, means and standard deviations of the 36 ten-day city-level SWIs computed separately over the years t , $(t-1)$ and $(t-2)$.

In addition, for every year t and every city, we compute and concatenate the mean SWI of all ten-day periods from April to September for (a) year t alone, (b) years t and $(t-1)$, (c) years t , $(t-1)$ and $(t-2)$. The period April to September corresponds to the dry season, as opposed to the period October to March, which corresponds to the wet season.

Moreover, for each quarter $1 \leq q \leq 4$ (January-March, April-June, July-September, October-December), for every year τ between 1959 and 2017 and every city α , we compute the average city-level SWI, denoted by $\overline{\text{SWI}}_{q,\tau,\alpha}$, and form the four cumulative distribution functions \hat{F}_q associated to the four data sets $\{\overline{\text{SWI}}_{q,\tau,\alpha} : 1959 \leq \tau \leq 2009, \alpha\}$. Then, for every year $1990 \leq t \leq 2017$ and every city α , we also add the 3×4 probabilities $\hat{F}_q(\overline{\text{SWI}}_{q,t,\alpha})$, $\hat{F}_q(\overline{\text{SWI}}_{q,t-1,\alpha})$, $\hat{F}_q(\overline{\text{SWI}}_{q,t-2,\alpha})$ ($q = 1, \dots, 4$). Here the years 1959 and 2017 correspond to the first year for which SWI data are available (1959) and to the last year considered in our study. The range 1959-2009 was the period considered by the Commission Interministérielle Catastrophe Naturelle to determine whether or not a city

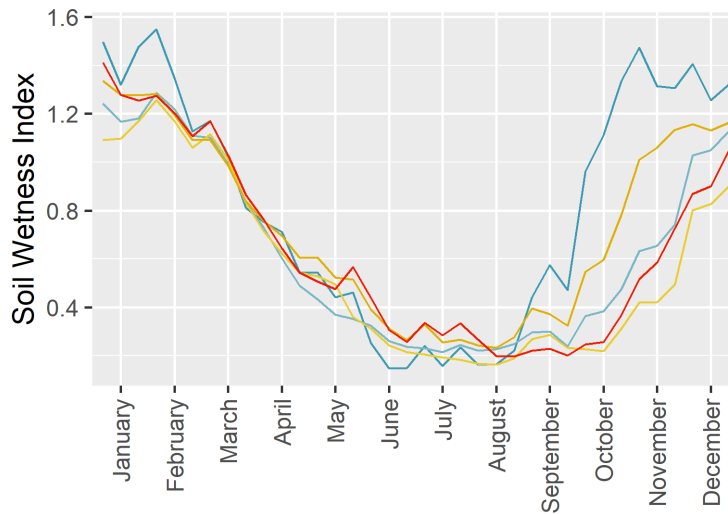


Figure 2.2: Chunks from five arbitrarily chosen time series of Soil Wetness Index (SWI) over the course of one year. It does not come as a surprise that the soil is drier during summer than during winter.

is eligible to obtain the government declaration of natural disaster for a drought event. The year 1990 is the first year for which we have data. These probabilities collectively provide a fine-grained description of the distribution of SWI.

2.2.3.3 On the city-level description

For every year, each city is described by a collection of covariates. A city's multi-faceted description attempts to capture all the city's traits that, beyond the city-level SWIs presented in the previous subsection, can explain the cost of a possible drought event. The collection should encompass covariates describing the *land use*, *clay concentration* and *physical and climatic profile*. *Land use* should be described in volume (including the number of houses, for instance), because larger and/or more populated cities incur larger costs, and in nature (including, for instance, the age of the housing stock, used as a proxy for the house building technology), because some buildings are more vulnerable than others [France Assureurs, 2022, page 28]. Including *clay concentration* is mandatory since it is the clay present in the soil that, by shrinking and swelling in dry and humid conditions, creates instabilities and generates cracks in buildings. Adding elements of a *physical and climatic* description is also relevant because, for instance, the ground slope and presence of trees near buildings influence the clay shrinking and swelling [Satriani et al., 2010, Zumrawi et al., 2017, Devkota et al., 2022]. Moreover, since the aftermath of a drought event can occur years after the drought event itself [Logar and van den Bergh, 2011, page 10, second paragraph], we include lagged values for the time-dependent covariates and past city-level costs.

The city-level description contains:

1. The year t .
2. The city's area, number of inhabitants, (estimated) number of houses located within the city's limits, house density, defined as the ratio of the number of houses to the city's area, and proportions of buildings built prior to 1949, between 1950 and 1974, between 1975 and 1989, and after 1989. The numbers of inhabitants come from [Antunez, 2022] (an R package that integrates data from the [Code Officiel Géographique de l'Insee](#)). The number of houses are estimated by CCR based on census data [Insee, 2000]. They are confidential. The city's areas are found in [IGN, 2018, Section 6.1, page 12, variable SUPERFICIE]. As for the proportions of buildings, they are computed based on data compiled by Insee and documented in [Insee, 2000]. Note that the thresholds 1949, 1974 and 1989 used to describe the age of the housing stock were fixed by Insee.
3. The (estimated) insured sum corresponding to the houses located within the city's limits, and the average house value, defined as the ratio of the aforementioned insured sum to the number of houses. The insured sums are evaluated by CCR based on data from Insee and portfolios data provided by CCR's cedents. The insured sums are confidential. We use the [Indice de la Fédération Française du Bâtiment](#) to account for inflation.
4. The proportions of the houses located within the city's limits that fall in each of the four clay shrinkage-swelling hazards categories. The four-category clay shrinkage-swelling hazard variable is defined, and obtained from, BRGM [MI, 2019]. Its resolution is fine enough for a city to fall in more than one category.
5. The city's average altitude, climatic zone, seismic zone, and proportions of surface with a "tree-coverage" greater than 10%. Here, climatic zone is a five-category variable [attributed by the French State](#) to each French department (one of the three levels of government under the national level, between the administrative regions and the communes; metropolitan France counts 96 departments). Seismic zone is a four-category variable attributed to each city by the French [Code de l'environnement](#). As for "tree-coverage", obtained from IGN, it corresponds to any of the 17 types of terrain documented in [IGN, 2021, Section 11.4, page 183, variable NATURE].
6. Six indicators of whether or not the city made a request for the government declaration of natural disaster for a drought event on years t to $(t-5)$, and six indicators of whether or not the city obtained the government declaration of natural disaster for a drought event on years t to $(t-5)$. We could derive these indicators because, being the secretary of the Commission Interministérielle Catastrophe Naturelle, CCR has access, every year, to the list of all cities that made a request for (and, possibly, obtained) the government declaration of natural disaster for a drought event. The data are publicly available [CCR, 2022a].

In addition, the city's description contains:

7. the cumulated city-level costs computed across the years $(t-1)$ to $(t-5)$;

8. the mean and median city-level costs of the drought events computed across the years $(t - 1)$ to $(t - 5)$ and all cities within the same department.

The city’s description is finally enriched with *compound covariates*. The compound covariates have a similar form. For every year t and each city α , for a given covariate $C_{h,\alpha,t}$ defined for all houses h within the city’s limits (and in the portfolios data provided by the cedents), we compute the weighted mean $\sum_h s_{h,\alpha,t} \times C_{h,\alpha,t} / \sum_h s_{h,\alpha,t}$ where $s_{h,\alpha,t}$ is the (estimated) insured sum of house h located within α ’s limits on year t . Here $C_{h,\alpha,t}$ can be:

9. the mean of all the t -specific 36 ten-day SWIs of the $8 \times 8 \text{ km}^2$ square which contains house h ;
10. the level of the clay shrinkage-swelling hazard localized at h (does not depend on t);
11. the ground slope localized at h (does not depend on t);
12. the three products $(9) \times (10)$, $(9) \times (11)$, $(9) \times (10) \times (11)$.

Moreover, for every year t and every city α , for each $C_{h,\alpha,t}$ among (9), (10) and (11), we also add the 30 29-quantiles of the data set $\{s_{h,\alpha,t} \times C_{h,\alpha,t} : h\}$ (where h ranges over the set of houses h within α ’s limits). The quantiles collectively provide a fine-grained description of the distributions of the covariates $\{s_{h,\alpha,t} \times C_{h,\alpha,t} : h\}$ where h ranges over the set of houses h within α ’s limits. Overall, the city’s description consists of a slightly fewer than 400 covariates.

2.3 The One-Step Ahead Sequential Super Learner

The One-Step Ahead Sequential Super Learner (OSASSL) adapts the canonical Super Learning methodology, one among many strategies to aggregate the predictions of several predictors. In Section 2.3.1, we give a brief historical perspective of the concept of aggregation and describe the canonical Super Learning methodology in the simple context where one learns from independent and identically distributed data. In Sections 2.3.2 and 2.3.3, we present an OSASSL built to forecast the cost of drought events, succinctly review its theoretical performance, and explain how it is used to forecast the cost of drought events. The complete theoretical analysis is developed in the Appendix.

2.3.1 Aggregation strategies

The idea of aggregating several estimation strategies to take advantage of their respective strengths emerged in the 1990s. The principle of “stacked generalization” was introduced by Wolpert [1992b]. Stacked generalization consisted in combining several lower-level predictive algorithms into a higher-level meta-algorithm with the aim of increasing predictive accuracy. Later, Breiman [1996b] showed how “stacking” can be used to improve predictive accuracy in a regression context, and how to impose constraints on the higher-level algorithm in order to achieve better predictive performance. Since then, stacking

has been evolving into a variety of methods among which is the canonical Super Learning methodology [van der Laan et al., 2007, Polley et al., 2011].

Related literature introduces and discusses the concepts of boosting, bagging, random forests [Freund, 1995, Breiman, 1996a, Amit and Geman, 1997, Breiman, 2001], and robust online aggregation (also known as prediction of individual sequences or prediction with expert advice) [Littlestone and Warmuth, 1994, Cesa-Bianchi and Lugosi, 2006]. A Bayesian perspective on “model averaging” was concomitantly introduced by Hoeting et al. [1999]. All these aggregation strategies have thrived both theoretically and in applications. In Section 2.4.3, we use several algorithms from the robust online aggregation literature as benchmarks.

Concisely, the canonical Super Learning methodology is a general methodology to learn a feature of the law of the data identified through an *ad hoc* risk function by relying on a library of (low-level) algorithms. The algorithms either compete (discrete Super Learning methodology) or collaborate through a (higher-level) meta-algorithm (continuous Super Learning methodology), with a cross-validation scheme determining the best performing algorithm or combination of algorithms, respectively. We refer the reader to [Naimi and Balzer, 2018] for a gentle introduction to Super Learning and a step-by-step development of two examples to illustrate concepts and address common concerns.

In the simpler case where one learns from independent and identically distributed data, one often implements a V -fold cross-validation scheme: first, the data set is split into V groups of roughly equal sizes (the “folds”); second, every algorithm is trained and tested V times, once for each fold, with each fold being used for testing after the algorithm has been trained using all the other folds; third, the cross-validated (empirical) risk of the algorithm is defined as the average of the V fold-specific (empirical) risks obtained by testing. In the present study, however, we learn from a (short) time series (with time-specific observations consisting of many dependent data-structures) and thus cannot rely on a V -fold cross-validation scheme. Instead, like Benkeser et al. [2018], we rely on a sequential cross-validation scheme: sequentially at each time t , for each algorithm: all data till time $(t - 1)$ are used for training and the t -specific data are used for testing; the t -specific cross-validated (empirical) cumulative risk of the algorithm is defined as the average of the τ -specific (empirical) risks (where τ ranges between 1 and t) obtained by testing. Remarkably, the sequential cross-validation scheme can neglect the dependence structure within each time-specific observation (in particular, it is not necessary to cross-validate spatially).

2.3.2 Presentation and theoretical performance of an OSASSL built to forecast the cost of drought events

We present here an OSASSL specifically built to forecast the cost of drought events. It is an instance of the general OSASSL fully developed and studied in the Appendix.

We let $(\bar{O}_t)_{t \geq 1}$ denote the time series that formalizes the data described in Section 2.2. At each time $t \in \mathbb{N}^*$, \bar{O}_t consists of the finite collection $(O_{\alpha,t})_{\alpha \in \mathcal{A}}$ of the (α, t) -specific observations, where each $\alpha \in \mathcal{A}$ represents a French city. For every $\alpha \in \mathcal{A}$, $O_{\alpha,t}$ decomposes as $O_{\alpha,t} := (Z_{\alpha,t}, X_{\alpha,t}, W_{\alpha,t}, Y_{\alpha,t}) \in \mathcal{Z} \times \mathcal{X} \times \{0, 1\} \times [0, B] =: \mathcal{O}$ where

- $Y_{\alpha,t} \in [0, B]$ is the city-specific cost of the drought event that year (known to take its values between 0 and a constant B , see Section 2.2.3.1),
- $W_{\alpha,t} \in \{0, 1\}$ is an indicator of whether or not the city obtained the government declaration of natural disaster for a drought event on year t ,
- $X_{\alpha,t} \in \mathcal{X}$ is the collection of covariates describing the city α on year t beyond $W_{\alpha,t}$ (see Section 2.2.3.3),
- $Z_{\alpha,t} \in \mathcal{Z}$ is the city-level SWI describing the drought event that year (see Section 2.2.3.2).

By convention, $W_{\alpha,t} = 0$ if city α did not obtain the government declaration of natural disaster for a drought event on year t and, in that case, $Y_{\alpha,t} = 0$.

For every $t \geq 1$, we let F_{t-1} be the σ -field $\sigma(O_{\alpha,\tau} : \alpha \in \mathcal{A}, 1 \leq \tau < t)$ generated at time t by past observations (by convention, $F_0 := \emptyset$). In view of assumption **A3** in Section 2.6, we assume that, for all $\alpha \in \mathcal{A}$ and $t \geq 1$, the conditional law of $Y_{\alpha,t}$ given $(W_{\alpha,t}, X_{\alpha,t}, Z_{\alpha,t}) = (1, x, z)$, $(Z_{\alpha',t})_{\alpha' \in \mathcal{A}}$ and F_{t-1} admits a conditional density $y \mapsto f^*(y|x, z)$ with respect to some measure on $[0, B]$. In words, this stationarity assumption states that, for all $\alpha \in \mathcal{A}$ and $t \geq 1$, the above conditional law of $Y_{\alpha,t}$ depends on (x, z) but neither on past observations (*i.e.*, F_{t-1}) nor on the city-level SWI of other cities (*i.e.*, $(Z_{\alpha',t})_{\alpha' \neq \alpha}$) nor on (α, t) . Therefore, the conditional expectation $y \mapsto \theta^*(y|x, z)$ of $Y_{\alpha,t}$ given $(W_{\alpha,t}, X_{\alpha,t}, Z_{\alpha,t}) = (1, x, z)$ (for all (x, z) in the support of any $(X_{\alpha,t}, Z_{\alpha,t})$ conditionally on $W_{\alpha,t} = 1$) becomes an eligible feature of interest of the law of $(\bar{O}_t)_{t \geq 1}$.

Under this stationarity assumption, we can use the estimator of the mean conditional cost to make predictions at any (x, z) provided that (x, z) falls in the domain of the observed $(X_{\alpha,t}, Z_{\alpha,t})$. Naturally, the scarcer the available information around (x, z) , the less reliable the prediction. Moreover, if (x, z) falls outside the domain, then, although a prediction may be made nonetheless, it should be taken with a grain of salt. So, in view of climate change, not-too-distant-future projections of drought events can be made.

The OSASSL built to forecast the cost of drought events is a meta-algorithm that learns the mean conditional cost θ^* from $(\bar{O}_t)_{t \geq 1}$ by stacking the estimators of θ^* provided by a user-supplied collection of J algorithms $\hat{\theta}_1, \dots, \hat{\theta}_J$. At each time $t \geq 1$, every algorithm $\hat{\theta}_j$ trained on $\bar{O}_1, \dots, \bar{O}_t$ outputs an estimator $\theta_{j,t}$ of θ^* . The OSASSL selects the best algorithm indexed by \hat{j}_t defined as the minimizer of the empirical average cumulative risks,

$$\hat{j}_t \in \arg \min_{1 \leq j \leq J} \hat{R}_{j,t}, \quad (2.1)$$

where

$$\hat{R}_{j,t} := \frac{1}{t|\mathcal{A}|} \sum_{\tau=1}^t \sum_{\alpha \in \mathcal{A}} [Y_{\alpha,\tau} - \theta_{j,\tau-1}(X_{\alpha,\tau}, Z_{\alpha,\tau})]^2 \mathbf{1}\{W_{\alpha,\tau} = 1\}. \quad (2.2)$$

Interestingly, the OSASSL is an online algorithm if each of the J algorithms $\hat{\theta}_1, \dots, \hat{\theta}_J$ is online, that is, such that the making of $\theta_{j,t}$ consists in an update of $\theta_{j,t-1}$ based on newly accrued data \bar{O}_t .

The t -specific measure of performance of each $\hat{\theta}_j$ is the unknown quantity

$$\tilde{R}_{j,t} := \frac{1}{t|\mathcal{A}|} \sum_{\tau=1}^t \sum_{\alpha \in \mathcal{A}} \mathbb{E} \left\{ [Y_{\alpha,\tau} - \theta_{j,\tau-1}(X_{\alpha,\tau}, Z_{\alpha,\tau})]^2 \mathbf{1}\{W_{\alpha,\tau} = 1\} \middle| \bar{Z}_\tau, F_{\tau-1} \right\}. \quad (2.3)$$

It takes the form of an average cumulative risk conditioned on $\bar{Z}_1 := (Z_{\alpha,1})_{\alpha \in \mathcal{A}}, \dots, \bar{Z}_t := (Z_{\alpha,t})_{\alpha \in \mathcal{A}}$ and on F_{t-1} . The t -specific oracular meta-algorithm is indexed by the oracular \tilde{j}_t defined as the minimizer

$$\tilde{j}_t \in \arg \min_{1 \leq j \leq J} \tilde{R}_{j,t}, \quad (2.4)$$

which, like each $\tilde{R}_{j,t}$, is unknown to us. Note that $\hat{R}_{j,t}$ estimates $\tilde{R}_{j,t}$ and that (2.1) mimics (2.4). By analogy, we also introduce

$$\tilde{R}_t^* := \frac{1}{t|\mathcal{A}|} \sum_{\tau=1}^t \sum_{\alpha \in \mathcal{A}} \mathbb{E} \left\{ [Y_{\alpha,\tau} - \theta^*(X_{\alpha,\tau}, Z_{\alpha,\tau})]^2 \mathbf{1}\{W_{\alpha,\tau} = 1\} \middle| \bar{Z}_\tau, F_{\tau-1} \right\}, \quad (2.5)$$

which can be interpreted as the average cumulative risk till time $t \geq 1$ of a dummy oracular algorithm that constantly maps $\bar{O}_1, \dots, \bar{O}_t$ to θ^* (the algorithm is said dummy because it does not learn).

The theoretical analysis hinges on a key-assumption about the dependence structure in the time series $(\bar{O}_t)_{t \geq 1}$, see assumption **A1** in Section 2.6 which uses conditional dependency graphs to model the amount of conditional independence. Specifically, we assume the existence of a graph \mathcal{G} with vertex and edge sets \mathcal{A} and \mathcal{E} such that if $\alpha \in \mathcal{A}$ is not connected by any edge $e \in \mathcal{E}$ to any $\alpha' \in \mathcal{A}' \subset \mathcal{A}$, then $O_{\alpha,t}$ is conditionally independent of $(O_{\alpha',t})_{\alpha' \in \mathcal{A}'}$ given F_{t-1} and \bar{Z}_t . Then what matters is the connectedness of the graph, as reflected by its degree, $\deg(\mathcal{G})$, which equals 1 plus the largest number of edges that are incident to a vertex in \mathcal{G} .

We emphasize that the dependency graph \mathcal{G} plays no role in the OSASSL's characterization and training. In other words, we can altogether neglect the intricate spatial dependence within each \bar{O}_t . However, the key-assumption is pivotal in the algorithm's theoretical analysis.

The performance of \hat{j}_t as an estimator of \tilde{j}_t is expressed in terms of a comparison of the excess average cumulative risk of the former to the excess average cumulative risk of the latter, using \tilde{R}_t^* (2.5) as a reference. In view of Corollary 2 in Section 2.6, if assumptions **A1** and **A2** hold (its additional assumptions **A3**, **A4**, **A5** are met, as shown in Section 2.7), then there exists a decreasing function $C : \mathbb{R}_+^* \rightarrow \mathbb{R}_+^*$ such that, for any $\varepsilon > 0$,

$$\mathbb{E} \left[\underbrace{\tilde{R}_{\hat{j}_t,t} - \tilde{R}_t^*}_{\text{excess risk of } \hat{j}_t} - (1 + \varepsilon) \left(\underbrace{\tilde{R}_{\tilde{j}_t,t} - \tilde{R}_t^*}_{\text{excess risk of } \tilde{j}_t} \right) \right] \leq C(\varepsilon) \frac{\log(J \log(\mathcal{I}^2))}{\mathcal{I}^2} \quad (2.6)$$

min.	1st qu.	median	mean	3rd qu.	99%-qu.	max
0	5	6	5.96	7	11	29

Table 2.1: Quartiles, 99%-quantile and mean of the numbers of neighboring cities in France in 2019. Although the maximum cannot be interpreted literally as $\deg(\mathcal{G}) - 1$, it nevertheless gives a sense of what a meaningful value of $\deg(\mathcal{G})$ could be.

where \mathcal{I} grows like the amount of information available and can be equal to either \sqrt{t} or $\sqrt{|\mathcal{A}|/(t \deg(\mathcal{G}))}$. The above oracular inequality follows from arguments that either ignore or exploit the conditional dependency graph \mathcal{G} used to model the amount of conditional independence. It has a familiar flavor for whoever is interested in Super Learning or, more generally, the aggregation or stacking of algorithms. In particular, the number J of algorithms plays a role in the error term only through its logarithm.

If the ratio $|\mathcal{A}|/\deg(\mathcal{G})$ is sufficiently large (both in absolute terms and relative to t), then the oracular inequality (2.6) is sharper when $\mathcal{I}^2 = |\mathcal{A}|/(t \deg(\mathcal{G}))$ than when $\mathcal{I}^2 = t$. This reveals that the OSASSL can leverage a large ratio $|\mathcal{A}|/\deg(\mathcal{G})$ in the face of a small t . This aspect is further discussed in the Appendix.

We conclude this section with some comments on \mathcal{G} , its degree and t . The dependency graph \mathcal{G} used to model the amount of conditional independence in **A1** operationalizes two different types of spatial dependence: one *geographical* and the other *administrative*. The former corresponds to the dependence caused by the proximity between two cities in geological and meteorological terms as well as in terms of vegetation. The latter corresponds to the dependence caused by the proximity between two cities. This is especially relevant when two cities belong to a same “communauté de communes” (*i.e.*, community of communes, a federation of municipalities). This second type of spatial dependence is less obvious than the first one. It arises from the fact that a declaration of natural disaster must be requested by the mayor of a city (see Section 2.1). If, say in a small federation, a mayor makes such a request, then it is more likely that the other mayors will as well. In the application, $t \approx 25$, $|\mathcal{A}| \approx 36,000$. As for $\deg(\mathcal{G})$, it is much harder to assess a meaningful value. In this regard, it is relevant to recall that, in 2019, France had around 1,000 federations of cities, each regrouping 30 cities on average. We computed the number of neighboring cities for each city. The quantiles and mean of these numbers are reported in Table 2.1. In particular, the city with the largest number of neighboring cities (Paris) has 29 of them.

2.3.3 Forecasting the cost of drought events

The OSASSL presented in Section 2.3.2 is designed to learn the mean conditional cost θ^* from $(\bar{O}_t)_{t \geq 1}$. At each time $t \geq 1$, it outputs the t -specific estimator $\hat{\theta}_{j_t, t}$. This estimator can be evaluated at every $(X_{\alpha, t+1}, Z_{\alpha, t+1})$ ($\alpha \in \mathcal{A}$) and we use the sum

$$\sum_{\alpha \in \mathcal{A}} \hat{\theta}_{j_t, t}(X_{\alpha, t+1}, Z_{\alpha, t+1}) \mathbf{1}\{W_{\alpha, t+1} = 1\}$$

to predict the cost of the drought event at time $(t + 1)$, that is, $\sum_{\alpha \in \mathcal{A}} Y_{\alpha, t+1} \mathbf{1}\{W_{\alpha, t+1} = 1\} = \sum_{\alpha \in \mathcal{A}} Y_{\alpha, t+1}$.

2.4 Application

This section discusses the practical implementation, training and exploitation of the OSASSL presented and studied in Section 2.3. Section 2.4.1 describes the collection of J algorithms $\hat{\theta}_1, \dots, \hat{\theta}_J$. Section 2.4.2 explains how the OSASSL is trained. Section 2.4.3 presents the results and comments upon them.

2.4.1 Implementing two OSASSLs

We deploy two meta-algorithms taking the form of OSASSLs, the discrete and continuous overarching Super Learners. Both rely on the same library of J algorithms $\hat{\theta}_1, \dots, \hat{\theta}_J$. These J algorithms are themselves OSASSLs either in the strict or in a loose sense (more details to follow).

2.4.1.1 Penalization

Because our ultimate goal is to forecast the cost of the latest drought event, we made the decision to rely on a penalized version of $\hat{R}_{j,t}$ (2.10), by substituting

$$\hat{R}_{j,t} + \frac{0.05}{t} \sum_{\tau=1}^t \left(\underbrace{\sum_{\alpha \in \mathcal{A}} Y_{\alpha, \tau} \mathbf{1}\{W_{\alpha, \tau} = 1\}}_{\text{actual cost}} - \underbrace{\sum_{\alpha \in \mathcal{A}} \theta_{\hat{j}_{\tau-1, \tau-1}}(X_{\alpha, \tau}, Z_{\alpha, \tau}) \mathbf{1}\{W_{\alpha, \tau} = 1\}}_{\text{predicted cost}} \right)^2 \quad (2.7)$$

for $\hat{R}_{j,t}$ (we recall that $\theta_{j,t}$ is the output of $\hat{\theta}_j$ trained on $\bar{O}_1, \dots, \bar{O}_t$ and that \hat{j}_t is defined in (2.1)). Observe that each t -specific penalization term equals 0.05 times the average over $1 \leq \tau \leq t$ of the τ -specific squared difference between the actual cost of the drought event (left-hand side, LHS, summand) and the predicted cost made by the (penalized) OSASSL trained on $\bar{O}_1, \dots, \bar{O}_{\tau-1}$ (right-hand side, RHS, summand). The factor 0.05 was chosen somewhat arbitrarily.

By adding this penalization term, the OSASSL favors the algorithms that better predict not only the *city-specific* costs but also the *overall* cost of the next drought event. In addition, the penalization term slightly dilutes the importance of the city-specific costs and, on the contrary, reinforces the importance of the overall cost, the latter being more dependable than the former as we explained in Section 2.2.3.

2.4.1.2 The discrete and continuous overarching Super Learners

Called the *discrete* overarching Super Learner, the first OSASSL is the algorithm that, at time $t \geq 1$, outputs $\theta_{\hat{j}_{t,t}}$ (using (2.7) instead of (2.2) as an empirical measure of the risk). In words, at time $t \geq 1$, the algorithm whose penalized empirical average

cumulative risk is the smallest is determined and the discrete overarching Super Learner returns the output of that algorithm trained on all data till time t .

We also consider a second OSASSL which is defined as a regular OSASSL based on a library derived from $\widehat{\theta}_1, \dots, \widehat{\theta}_J$ and comprising $J' = \mathcal{O}(\varepsilon^{1-J})$ algorithms where $\varepsilon > 0$ is a small positive number ($J' = \mathcal{O}(\varepsilon^{1-J})$ means that J' is upper-bounded by a constant times ε^{1-J}). Specifically, these J' algorithms are denoted by $\widehat{\theta}_\pi$ where the index π ranges in an ε -net over the simplex $\{x \in (\mathbb{R}_+)^J : \sum_{j=1}^J x_j = 1\}$ (an ε -net whose cardinality is J' , that is, a finite subset of J' elements of the simplex which ‘‘approximates’’ the simplex). For each π in the ε -net, $\widehat{\theta}_\pi$ trained on $\bar{O}_1, \dots, \bar{O}_t$ outputs the π -specific convex combination $\sum_{j=1}^J \pi_j \theta_{j,t}$. The bound in (2.6) is still meaningful when $\varepsilon = \mathcal{O}(\mathcal{I}^{-1})$. We refer to this second OSASSL as the *continuous* overarching Super Learner.

2.4.1.3 The discrete and continuous overarching Super Learners’ library of algorithms

We now turn to the description of the J algorithms $\widehat{\theta}_1, \dots, \widehat{\theta}_J$. All of them rely on a collection of base learners $\widehat{\mathcal{L}}_1, \dots, \widehat{\mathcal{L}}_K$. Some of the base learners rely on linear models and their extensions (lasso, ridge, elastic net, multivariate adaptive regression splines, support vector regression). Others are tree-based algorithms (CART, random forest, gradient boosting), or rely on neural networks. Others fall in the category of k -nearest-neighbors algorithms tailored to our study so that the dissimilarity between observations is a convex combination of the Kolmogorov-Smirnov distances between the empirical average cumulative distribution functions mentioned in Section 2.2.3. Finally, some are regular Super Learners themselves, based on a selection of the aforementioned base learners and oblivious to the temporal ordering (that is, they rely on vanilla inner V -fold cross-validation).

Moreover, some of these base learners are combined (upstream) with screening algorithms. A screening algorithm is merely an algorithm that selects a subset of the covariates deemed relevant to feed the base learners. In general, the selection can be either deterministic or data-driven. In our study, we only use deterministic screening algorithms based on expert knowledge.

Overall, we implement a collection of $K = 74$ base learners (including the variants obtained by combining with different screening algorithms). The collection is shared by the J algorithms $\widehat{\theta}_1, \dots, \widehat{\theta}_J$ which differ in the methods they rely on to exploit the base learners.

One of the method yields a OSASSL precisely as defined in (2.1) and (2.2)/(2.7) where we substitute K for J and $\ell_{j,\tau-1}$ for $\theta_{j,\tau-1}$, with $\ell_{j,t}$ the output of $\widehat{\mathcal{L}}_j$ trained on $\bar{O}_1, \dots, \bar{O}_t$. The resulting OSASSL is an instance of discrete Super Learner as previously described when introducing the first overarching Super Learner. As we already explained, the library of base learners $\widehat{\mathcal{L}}_1, \dots, \widehat{\mathcal{L}}_K$ can be extended using an ε -net over the simplex $\{x \in (\mathbb{R}_+)^K : \sum_{k=1}^K x_k = 1\}$. For each π in the ε -net, $\widehat{\mathcal{L}}_\pi$ trained on $\bar{O}_1, \dots, \bar{O}_t$ outputs the π -specific convex combination $\sum_{k=1}^K \pi_k \ell_{k,t}$. Using the extended collection of base learners, the same method then yields an instance of continuous Super Learner as previously described when introducing the second overarching Super Learner.

In a similar fashion, we consider several methods to exploit the base learners $\widehat{\mathcal{L}}_1, \dots, \widehat{\mathcal{L}}_K$. Heuristically, the principle is to learn to produce a single prediction based on the multiple predictions made by the base learners once they have been trained, just like we described in the previous paragraph. Two natural and simple methods consist in using the average or the median of the base learners' predictions. Some methods rely on the same method as in the previous paragraph with an extra penalization term in the definition of the risk (similar to the one used to define (2.7) based on (2.2)). The other methods rely on the lasso, ridge and elastic net algorithms, or on the random forests, gradient boosting and support vector regression algorithms. Finally, some of the methods can exploit the covariates. Overall, we implement a collection $J = 50$ algorithms $\widehat{\theta}_1, \dots, \widehat{\theta}_J$.

2.4.2 Training the discrete and continuous overarching Super Learners

At each time $t \geq 1$ we define a summary of the past based on observations made during the five previous years. To do so, we reserve the data from year 1990 to year 1994. This is very relevant for two reasons. First, a drought-related claim can be the by-product of repeated shrinkage-swelling episodes over the years. Second, a city-level cost of a drought event is expected to be high when the city did not benefit recently from a government declaration of natural disaster for a drought event (because of the possible accumulation of damages over the years); on the contrary, it is expected to be low otherwise (because damages may already have been compensated).

For each $t \in \{1995, \dots, 1999\}$, we derive $\ell_{1,t-1994}, \dots, \ell_{K,t-1994}$. For each year $t \in \{2000, \dots, 2005\}$, we derive $\theta_{1,t-1994}, \dots, \theta_{J,t-1994}$ using $\ell_{1,(t-1)-1994}, \dots, \ell_{K,(t-1)-1994}$, and we also derive $\ell_{1,t-1994}, \dots, \ell_{K,t-1994}$. For each $t \in \{2006, \dots, 2017\}$, we derive the discrete overarching Super Learner \widehat{j}_{t-1994} using $\theta_{1,(t-1)-1994}, \dots, \theta_{J,(t-1)-1994}$ (which rely themselves on $\ell_{1,(t-2)-1994}, \dots, \ell_{K,(t-2)-1994}$), and we also derive $\theta_{1,t-1994}, \dots, \theta_{J,t-1994}$ and $\ell_{1,t-1994}, \dots, \ell_{K,t-1994}$. For each $t \in \{2006, \dots, 2017\}$, the continuous overarching Super Learner is derived too.

To this day, the real costs and city-level costs for the years 2018, 2019, 2020 and 2021 are still uncertain. We thus cannot train our algorithms beyond the year 2017.

Numerical analysis The numerical analysis was conducted in R [R Core Team, 2022]. We adapted the R package `SuperLearner` [Polley et al., 2021] in a package called `SequentialSuperLearner` [Chambaz and Ecoto, 2021]. To see the package in action, simply run in R the commands

- `library(SequentialSuperLearner)` and
- `example(overarching_SuperLearner)`.

Let us denote by $T_{k,1:t}^\ell$ (any $1 \leq k \leq K$ and $t \geq 1$) and $T_{j,2:t}^\theta$ (any $1 \leq j \leq J$ and $t \geq 2$) the time needed to train the base learner ℓ_k using $\bar{O}_1, \dots, \bar{O}_t$, thus obtaining $\ell_{k,t}$, and the time needed to train the algorithm θ_j using $\bar{O}_2, \dots, \bar{O}_t$ and the predictions $\ell_{k,\tau}(X_{\alpha,\tau+1}, Z_{\alpha,\tau+1})$ (all $1 \leq k \leq K$ and $1 \leq \tau < t$). Neglecting the time needed to make predictions and to determine the discrete and continuous overarching Super Learners,

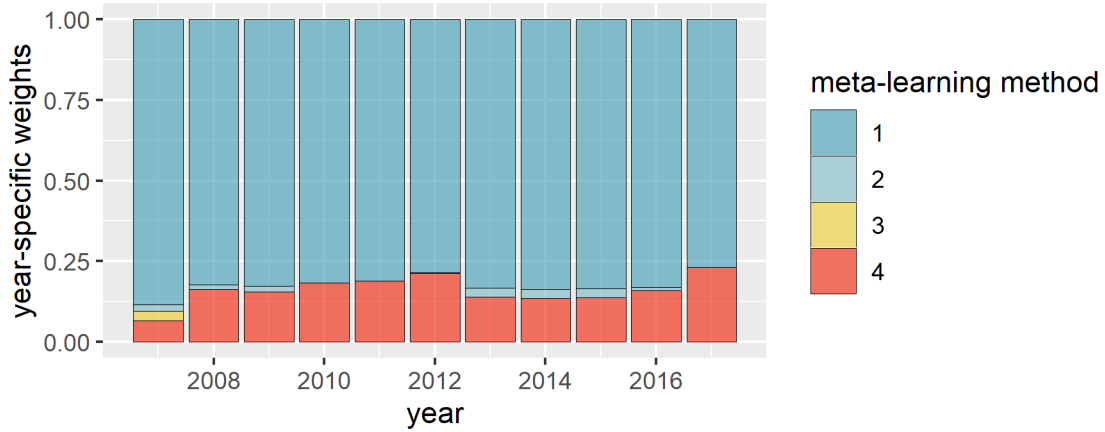


Figure 2.3: Evolution (from 2007 onward) of the weights attributed in the overarching Super Learner to four of the algorithms $\hat{\theta}_1, \dots, \hat{\theta}_J$. The others get no weight at all.

the computational time at time $t \geq 2$ is

$$\text{CompTime}(t) := \sum_{k=1}^K \sum_{j=1}^J \sum_{\tau=1}^{t-1} (T_{k,1:\tau}^{\ell} + T_{j,2:\tau+1}^{\theta}).$$

If we further assume that every $T_{k,1:t}^{\ell}$ and $T_{j,2:t}^{\theta}$ is upper bounded by κ times the number of (α, τ) -specific data-structure involved in the training, then

$$\text{CompTime}(t) \leq \kappa \times KJ \times |\mathcal{A}| \times (t-1)t.$$

Anecdotally, producing the results analyzed in the next section took a dozen hours of computations.

2.4.3 Results

In Figure 2.3 we present the evolution of the weights that characterize the continuous overarching Super Learner through the years 2007 to 2017. The figure reveals that only four of the $J = 50$ algorithms $\hat{\theta}_1, \dots, \hat{\theta}_J$ get a positive weight, and that only two of them do in 2016 and 2017. Moreover, one of the algorithms dominates the others during the whole training. It does not come as a surprise that this algorithm (whose method is a variant of gradient boosting with linear boosters) is constantly selected by the discrete overarching Super Learner.

For confidentiality reasons, we were not given the authorization to discuss how the overarching Super Learners fare compared to the algorithm currently deployed at CCR to predict the overall costs of drought events in France from 2007 to 2017. However, we were authorized to make a comparison for the sole year 2017. That particular year, the discrete and continuous overarching Super Learners outperform the algorithm currently

deployed at CCR, with a precision of 96% (discrete overarching Super Learner), 94% (continuous overarching Super Learners) versus 83% (currently deployed algorithm).

In Figure 2.4 we primarily present four sequences of predictions from 2007 to 2017: those from the discrete and continuous overarching Super Learners and those obtained by taking the average or the median of all the base learners' predictions. Secondly, we also summarize all the base learners' predictions with boxplots. The variability of the base learners' predictions is striking, confirming that the base learners can strongly disagree. Note that the two sequences of predictions from the Super Learners are quite similar. Overall, the Super Learners' predictions look generally accurate and better than the averaged predictions. As for the medians of the predictions, they seem to provide a better trade-off than the averages. However neither the method consisting in using the average of the base learners' predictions nor the method consisting in using their median is given a positive weight by the overarching Super Learner. Table 2.2 reports the averages and standard deviations (over the years) of the ratios of the predicted costs to the real costs for the predictors. Both in terms of mean and standard deviation, the discrete overarching Super Learner outperforms its continuous counterpart, which itself outperforms the predictors that average or take the median of all the base learners' predictions.

We also applied nine aggregation strategies from the robust online aggregation literature as benchmarks: online gradient descent [Zinkevich, 2003], exponentially weighted average aggregation, fixed-share aggregation, online ridge regression [for these three strategies, we refer to Cesa-Bianchi and Lugosi, 2006], follow the regularized leader [Shalev-Shwartz and Singer, 2007], MLewa, MLpol, MLprod [for these three strategies, we refer to Gaillard et al., 2014], and Bernstein online aggregation [Wintenberger, 2017]. We used the default implementations provided by the R package `opera` [Gaillard et al., 2023]. In each case, the predictions made by the base learners $\ell_{k,t-1994}$ ($1 \leq k \leq K$) are sequentially aggregated for t ranging from 2000 to 2017. Table 2.2 also reports the averages and standard deviations (over the years 2007 to 2017) of the ratios of the predicted costs to the real costs for the nine additional predictors thus derived. In terms of mean, the discrete overarching Super Learner and the predictions obtained by applying the exponentially weighted average aggregation or MLpol strategies are essentially on par. In terms of standard deviation, the discrete overarching Super Learner outperforms both strategies (as a matter of fact, all of them).

Furthermore, the two Super Learners' predictions are quite good for all years except 2012 and 2016. The poorer predictions in 2016 are more problematic because the real cost in 2016 is much higher than in 2012.

The year 2016 is known in the French insurance market as particularly challenging. Unfortunately, as far as we know, this fact is undocumented in the literature. However, we can report two facts to uphold this statement.

First, the year-specific average cost is particularly large in 2016 compared to the global average cost between 2007 and 2017: 797,000 euros versus 482,000 euros. By year-specific average cost we mean the ratio of the total cost of the year's drought event to the corresponding number of government declarations of natural disaster for a drought

predictions	mean	standard deviation
online ridge regression	0.009	0.004
exponentially weighted average aggregation	1.042	0.421
discrete overarching Super Learner	1.045	0.280
MLpol	1.047	0.394
MLewa	1.081	0.395
continuous overarching Super Learner	1.100	0.320
fixed-share aggregation	1.149	0.453
MLprod	1.175	0.403
follow the regularized leader	1.204	0.401
online gradient descent	1.207	0.423
median of the base learners' predictions	1.209	0.446
average of the base learners' predictions	1.211	0.420
Bernstein online aggregation	1.219	0.416

Table 2.2: Averages and standard deviations (over the years) of the ratios of the predicted costs to the real costs. The predictions are those made by the discrete and continuous overarching Super Learners or derived from all the base learners' predictions by averaging, taking their median, or employing nine different aggregation strategies from the robust online aggregation literature (see main text for details).

event delivered that year. By global average cost we mean the ratio of the total cost of the drought events between 2007 and 2017 to the total number of government declarations of natural disaster for a drought event delivered these years.

Second, we can quote Charpentier et al. [2022, end of Section 4.1] who say of their predictions for the year 2016 that they are “severely underestimated”. Judging by their Figure 7, the underestimation by the discrete and continuous overarching Super Learners for the year 2016 is less pronounced than the underestimation by their algorithms (but we recall that they tackle a more challenging problem than us because we focus on the city-specific costs for those cities that have obtained the government declaration of natural disaster for a drought event whereas they consider all French cities).

In Figure 2.5 we present (Gaussian) kernel density estimates of the conditional laws of the residual error (defined as the real cost minus the prediction made by the continuous overarching Super Learner – the figure is very similar when substituting the discrete overarching Super Learner for the continuous one) in ten strata characterized by the deciles of the city-level costs. We note that the higher the city-level costs, the higher the residuals. Moreover, the overarching Super Learner tends to overestimate the costs in cities with lower city-level costs and, on the contrary, it tends to underestimate them in cities with higher city-level costs.

In Figure 2.6 we present two maps that provide insight into the geographical distribution of the residual errors (of the predictions made by the continuous overarching Super Learner – the maps are very similar when considering its discrete counterpart). In the

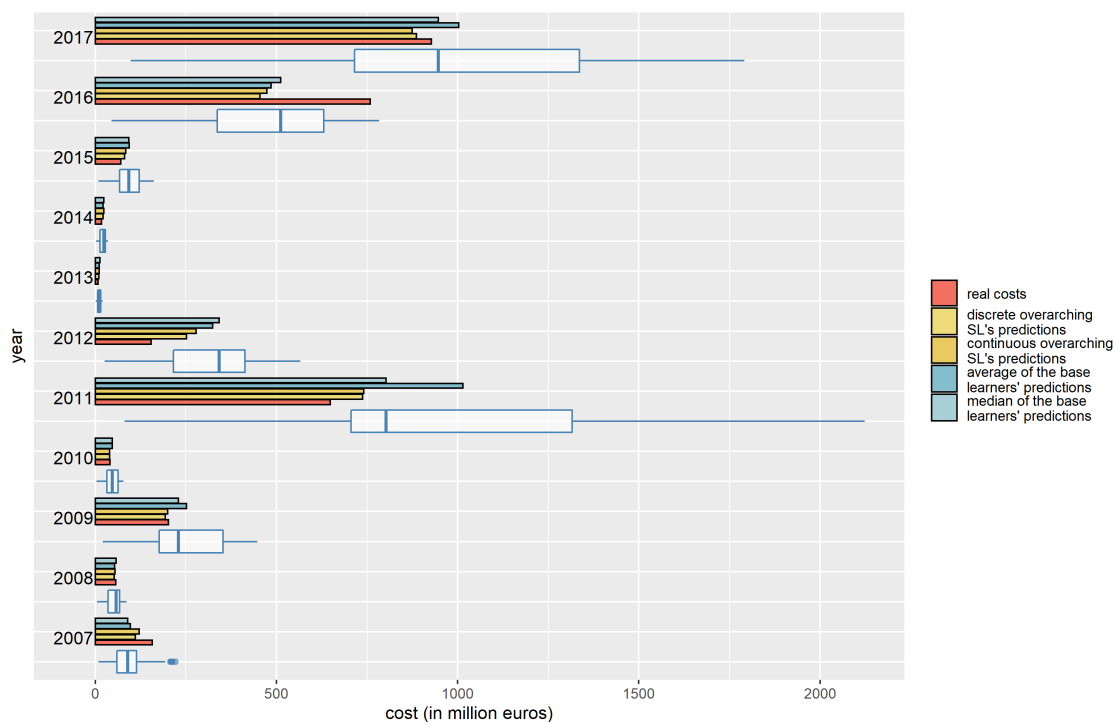


Figure 2.4: Presentation (from 2007 onward) of the real costs of drought events and their predictions. The predictions are either those made by the discrete (pale yellow) and continuous overarching (dark yellow) Super Learners or obtained by averaging all the base learners' predictions (red) or using their median (blue vertical bars). The figure also presents boxplots that summarize all the base learners' predictions. Note the high variability of these predictions. In this figure we use current euros.

LHS map, a city contributes as many points as the number of times it benefited from a government declaration of natural disaster for a drought event between 2007 and 2017. In the RHS map, a city contributes a point if and only if it benefited from a government declaration of natural disaster for a drought event in 2016, the year considered as particularly challenging. In both maps, the color reflects the quartile of the residual error to which the city- and time-specific residual error belongs. Moreover, in the LHS map the transparency reflects the number of times the city benefited from a government declaration of natural disaster for a drought event between 2007 and 2017, a larger number leading to less transparency. By comparing the two maps, we notice (i) that the 2016 drought episode impacted very strongly the South of France and (ii) that, in this region, the residual errors tend to be higher, leading to the underestimation of the local cost.

2.4.4 On the importance of the variables used to make predictions

The discrete and continuous overarching Super Learners make predictions based on a multi-faceted description of cities and their exposures consisting of slightly fewer than

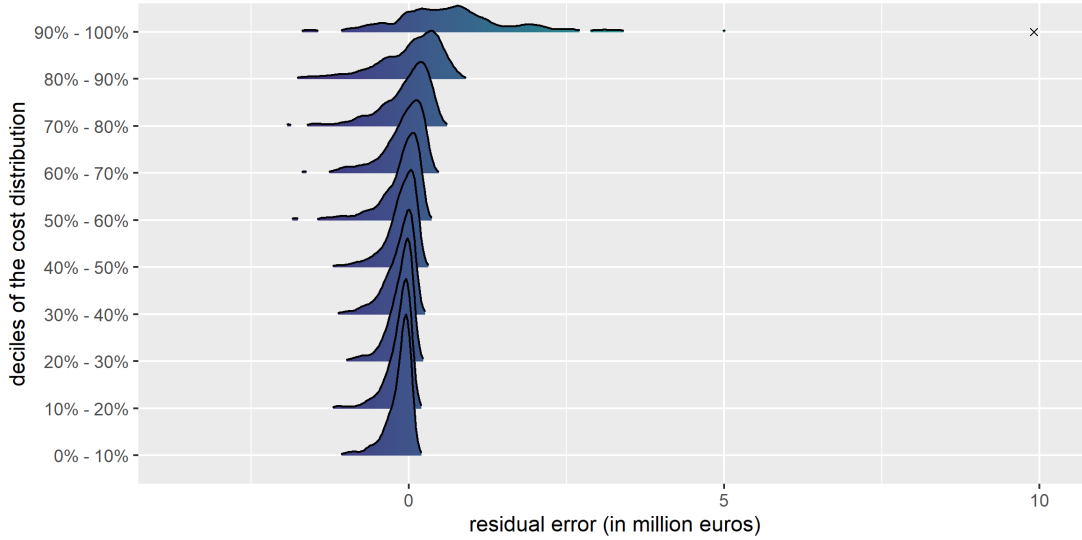


Figure 2.5: Kernel density estimates of the conditional laws of the residual error (of the predictions made by the continuous overarching Super Learner) in ten strata characterized by the deciles of the city-level costs. The cross at the upper RHS of the plot indicates the maximum residual error, made for a city belonging to the last decile of the cost distribution. In this figure we use current euros.

400 covariates (see Section 2.2). It is natural to wonder which variables mainly influence the prediction.

The question pertains to the definition and estimation of variable importance measures. The literature on this topic is rich and both [van der Laan, 2006, Hubbard et al., 2016, Williamson et al., 2021] on the one hand or [Lundberg and Lee, 2017, and references therein] on the other hand give insights about how to answer it. Unfortunately, building on these approaches is unrealistic because our data set consists of a *short* time series with time-specific observations consisting of *many dependent* data-structures and, to boot, because we are interested in a *high* number of covariates. We thus propose the following simple approach tailored to our needs.

For any time $t \geq 1$, the OSASSL outputs the t -specific estimator $\hat{\theta}_{j_t, t}$. Then, for every $\alpha \in \mathcal{A}$, evaluating this estimator at $(X_{\alpha, t+1}, Z_{\alpha, t+1})$ yields the prediction $\hat{Y}_{\alpha, t+1} := \hat{\theta}_{j_t, t}(X_{\alpha, t+1}, Z_{\alpha, t+1})\mathbf{1}\{W_{\alpha, t+1}\}$ of the cost $Y_{\alpha, t+1}$ of the drought event at time $(t+1)$ for city α . Highlighting that $X_{\alpha, t+1}$ and $Z_{\alpha, t+1}$ consist of (many) covariates, let us rewrite $(X_{\alpha, t+1}, Z_{\alpha, t+1}) =: (C_{s, \alpha, t+1} : 1 \leq s \leq S)$. Because $\hat{Y}_{\alpha, t+1} = 0$ if $W_{\alpha, t+1} = 0$ by design, we will not consider the importance of the covariate $W_{\alpha, t+1}$.

Fix $t = 2017$ and set arbitrarily $1 \leq s \leq S$. If s is such that the covariate $C_{s, \alpha, \tau}$ ($1 \leq \tau \leq t - 2006$) can be treated as a continuous variable, then we let ρ_s be the absolute value of the Spearman rank correlation coefficient [Hollander and Wolfe, 1999, Section 8.5] computed based on $((\hat{Y}_{\alpha, \tau}, C_{s, \alpha, \tau}) : \alpha \in \mathcal{A}, 1 \leq \tau \leq t - 2006)$. If s is such that $C_{s, \alpha, \tau}$ ($1 \leq \tau \leq t - 2006$) take v values (in which case $2 \leq v \leq 5$), then we let ρ_s be

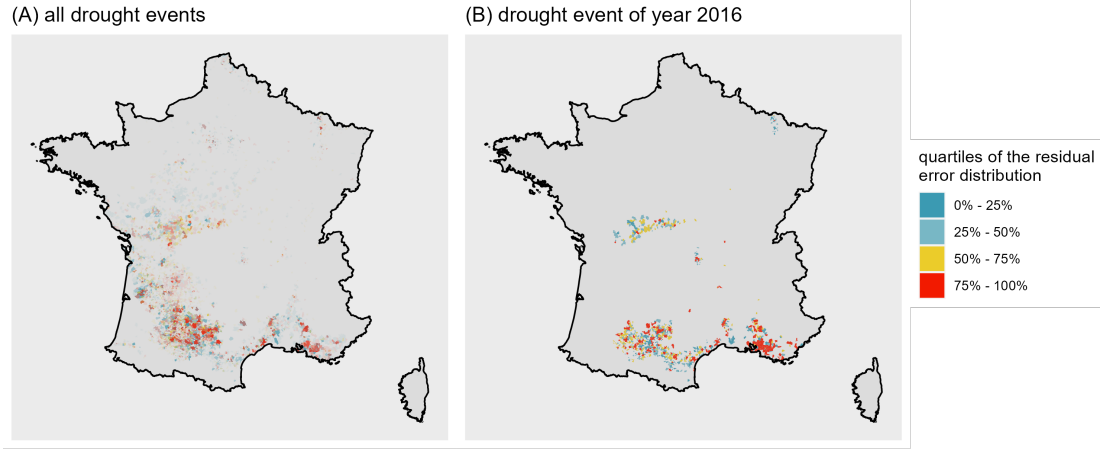


Figure 2.6: Geographical distribution of the residual errors (of the predictions made by the continuous overarching Super Learner). (A): a city contributes as many points as the number of times it benefited from a government declaration of natural disaster for a drought event between 2007 and 2017. (B): a city contributes a point if and only if it benefited from a government declaration of natural disaster for a drought event in 2016. The color reflects the quartile of the residual error to which the city- and time-specific residual error belongs (based on all the errors). In (A), the transparency reflects the number of times the city benefited from a government declaration of natural disaster for a drought event between 2007 and 2017, a larger number leading to less transparency.

the correlation ratio computed based on $((\hat{Y}_{\alpha,\tau}, C_{s,\alpha,\tau}) : \alpha \in \mathcal{A}, 1 \leq \tau \leq t - 2006)$:

$$\rho_s := \left(\frac{\sum_{\nu=1}^v n_\nu (\bar{y}_\nu - \bar{y})^2}{\sum_{\alpha \in \mathcal{A}} \sum_{\tau=1}^{t-2006} (\hat{Y}_{\alpha,\tau} - \bar{y})^2} \right)^{1/2}$$

where \bar{y}_ν is the average of the $\hat{Y}_{\alpha,\tau}$ s such that $C_{s,\alpha,\tau} = \nu$ and \bar{y} is the average of all $\hat{Y}_{\alpha,\tau}$ s. Note that we could have defined ρ_s as Wilcoxon test's statistic (case $v = 2$) or the Kruskal-Wallis test's statistics (case $3 \leq v \leq 5$) [see Hollander and Wolfe, 1999, Sections 3.1 and 6.1] but chose not to, preferring that all ρ_s s naturally lie in $[0, 1]$ to ease comparisons.

In all cases the larger is ρ_s the more we are willing to believe that the s -th covariate well explains the predictions made by the OSASSL. Note how we substituted the word “explains” for the word “influences”. This is an acknowledgement that our assessments simply rely on associations and have no causal interpretation. We resort to permutation tests to assess significance levels, with one million independent permutations drawn

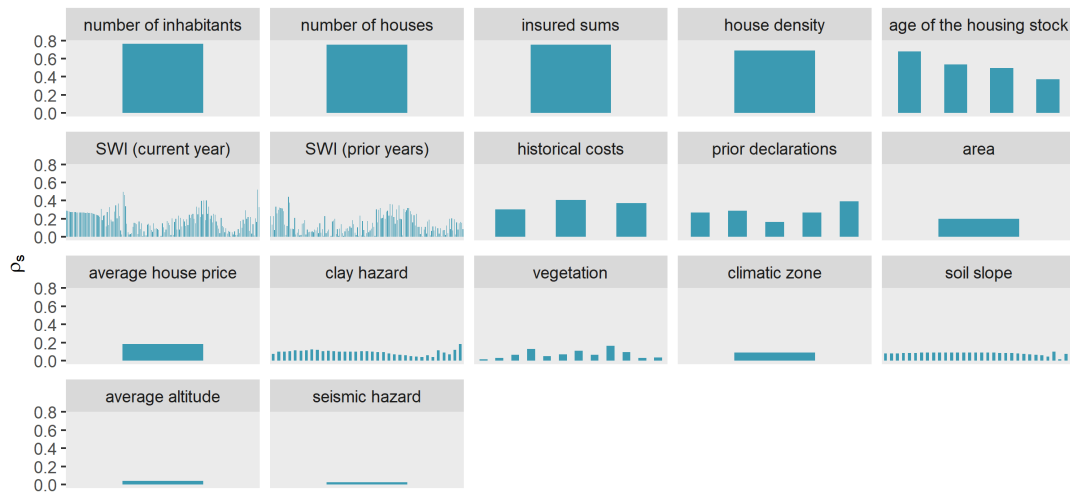


Figure 2.7: Assessing the importance of the variables used to make predictions by the discrete overarching Super Learner. The larger is ρ_s the more we are willing to believe that the s -th covariate well explains the predictions made by the discrete overarching Super Learner. Every ρ_s is declared significantly positive by a permutation test analysis.

uniformly in each of the above cases. For every $1 \leq s \leq S$, ρ_s is much larger (often by several orders of magnitude) than the maximum value obtained by permutation. This strongly supports the findings reported below.

Surprisingly, all the covariates are deemed important (based on the permutation tests), though some covariates are of course more important than others. We only report the values of $(\rho_s : 1 \leq s \leq S)$ for the discrete overarching Super Learner (those of the continuous overarching Super Learner are very similar). To do so, we regroup the covariates into 17 homogeneous categories: nine of them consist of a single covariate (the city's area, average altitude, average house price, climatic zone, house density, insured sums, number of inhabitants, number of houses, seismic hazard); four of them consist of a small number of covariates grouped by theme (related to the city's age of the housing stock, historical costs, prior government declarations of natural disaster for a drought event, vegetation); four of them gather many covariates by theme (related to the city's clay hazard, soil slope, SWI during the current year and history of SWI). Figure 2.7 summarizes the results. The most important variables (number of inhabitants, number of houses, insured sums) are related to a potential of exposure to drought events. The next two more important (group of) variables (house density and age of the housing stock) help to characterize the buildings at risk. The variables that we mentioned so far do not vary significantly over time. The next two more important (group of) variables (SWI, current and prior years) provide a meteorological description of the drought events, which obviously vary over time. As anticipated, relying on past descriptions of the drought events is relevant. The remaining (groups of) variables complete the characterization of the buildings at risk.

2.5 Discussion

The French legal framework known as the natural disasters compensation scheme was created in 1982. Drought events were included in 1989. Since then they have been the second most expensive type of natural disaster. In recent years, drought events have been remarkable in their extent and intensity. The problem is worsening and not limited to France, as was predicted in the technical report [Wüest et al., 2011, page 7]: “as our climate continues to change, the risk of property damage from soil subsidence [that is, drought events] is not only increasing but also spreading to new regions across Europe”.

Forecasting the cost of a drought event is an important actuarial problem. To tackle this challenge, we develop a new methodology that builds upon Super Learning, a popular aggregation strategy. Our overarching Super Learner blends predictions made by a collection of OSASSLs which, themselves, blend the predictions made by a variety of machine-learning algorithms.

The theoretical analysis hinges on a stationarity assumption stating that the mechanism producing a local drought-event-related cost conditionally on its local description remains constant throughout time and France. The assumption warrants both the possibility to define and estimate the mean conditional cost on the one hand and the use of its estimator to make predictions on the other hand. Predictions can be made at any local description (x, z) provided that it falls in the domain of the observed local descriptions. Naturally, the scarcer the available information around (x, z) , the less reliable the prediction. In addition, if (x, z) falls outside the domain then, although a prediction may be made nevertheless, it should be taken with a grain of salt. Therefore, in view of climate change, it is meaningful to make projections of drought events in the not-too-distant future. Under another assumption on the complex dependence structure induced in the data by the spatial and temporal nature of the phenomenon of drought, we showed that OSASSL can learn the mean conditional cost, making up for the shortness of the time series thanks to the manyness of each time-specific observation because the latter are only slightly dependent.

We present two implementations, called the discrete and continuous overarching Super Learners. Their predictions are generally accurate and better than those obtained, for instance, by averaging or taking the median of all the low-level predictions made by the base machine-learning algorithms (two ways among 50 implemented to combine the 74 low-level predictions). We also applied nine aggregation strategies from the robust online aggregation literature as benchmarks. The best performing strategies, the exponentially weighted average aggregation and MLpol strategies [Cesa-Bianchi and Lugosi, 2006, Gaillard et al., 2014], are on par with the discrete overarching Super Learner in terms of mean over the years of the ratios of the predicted costs to the real costs, and outperformed by it in terms of their standard deviation. Specifically, the two Super Learners’ predictions are quite good for all years except 2012 and 2016. The poorer predictions in 2016, a year known in the French insurance market to be particularly challenging, are more problematic because the real cost in 2016 is much higher than in 2012. Moreover, we were given the authorization to compare the predictions of the discrete and contin-

uous overarching Super Learners with that of the algorithm currently deployed at CCR for the sole year 2017: the precisions are respectively 96% (discrete overarching Super Learner), 94% (continuous overarching Super Learners) and 83% (currently deployed algorithm).

The quality of the predictions made by the overarching Super Learners strongly depends on the relevance and quality of the covariates used to make predictions — in particular, on the local description of the drought event. Regarding the covariates’ relevance, we develop an *ad hoc* approach to define and estimate variable importance measures so as to assess which covariates mainly influence the predictions. We acknowledge that the word “influence” is somewhat misleading because a causal interpretation is out of reach and our assessments simply rely on associations. Surprisingly, all of the covariates are deemed relevant (based on permutation tests), though some covariates are more important than others. Regarding the covariates’ quality, the overarching Super Learners would probably benefit from a refined version of the city-level SWI that, contrary to the one we rely on, does not assume that the nature of the soil is the same all over France. In addition, the local description would also be considerably enhanced if it included information such as the distribution of the proximity between a house and a tree at the city-level, or the distribution of the depth of house foundations at the city-level. The local description could also be enhanced by including direct measurements of soil shrinkage and swelling which can be obtained by radar interferometry.

In this study, we forecast the cost of drought events in France by Super Learning for those cities that have obtained the government declaration of natural disaster for a drought event. The next step will be to predict which cities will obtain the government declaration of natural disaster for a drought event. Tackling this difficult challenge will allow forecasting the cost of drought events earlier.

2.6 Appendix: the OSASSL and its oracular performances

In this section and the next, we develop and theoretically analyze an OSASSL to learn sequentially from a time series $(O_{\alpha,1})_{\alpha \in \mathcal{A}}, \dots, (O_{\alpha,t})_{\alpha \in \mathcal{A}}$ a feature of its law which does not depend on (α, t) . We do so in a general framework that encompasses the model described in Section 2.3.2, a model that we use to illustrate the assumptions. Two main arguments justify that approach. First, beyond the present study, learning from sequences of data is an important topic especially when each data-structure is a complex object with an intricate dependence pattern. Second, adopting a more generic viewpoint makes it easier to identify the key points and to exploit our findings in different settings. For compactness, we use from now on the notation $\llbracket n \rrbracket := \{1, \dots, n\}$ for any integer $n \geq 1$.

We rely on conditional dependency graphs to model the amount of conditional independence.

Assumption 1. There exists a graph \mathcal{G} with vertex set \mathcal{A} such that if $\alpha \in \mathcal{A}$ is not connected by any edge to any vertex in $\mathcal{A}' \subset \mathcal{A}$, then $O_{\alpha,t}$ is conditionally independent of

$(O_{\alpha',t})_{\alpha' \in \mathcal{A}'}$ given F_{t-1} and (possibly) a known, fixed summary measure $\bar{Z}_t := \text{Summ}(\bar{O}_t)$ of each observation \bar{O}_t .

As emphasized in Section 2.3.2, the dependency graph \mathcal{G} plays no role in the one-step ahead sequential Super Learner's characterization and training, but it plays a central role in the algorithm's theoretical analysis. For every $t \geq 1$ the summary measure \bar{Z}_t writes as $\bar{Z}_t := (Z_{\alpha,t})_{\alpha \in \mathcal{A}}$ where each $Z_{\alpha,t}$ belongs to a common set \mathcal{Z} . The summary measure is said *fixed* because it is derived from \bar{O}_t by evaluating at \bar{O}_t the fixed (in $t \geq 1$ and $\alpha \in \mathcal{A}$) function Summ . The adverb *possibly* hints at the case where Summ maps every \bar{O}_t to an uninformative, empty summary.

Motivating example of Section 2.3.2. Here, \mathcal{A} represents the set of French cities, $\mathcal{O} = \mathcal{Z} \times \mathcal{X} \times \{0, 1\} \times [0, B]$ and the function Summ in **A1** merely extracts from \bar{O}_t the collection of all the city-level SWI. \square

Our main objective is to estimate a feature θ^* of the law \mathbb{P} of $(\bar{O}_t)_{t \geq 1}$, an element of a parameter space Θ that is known to minimize over Θ the risk induced by a loss ℓ and \mathbb{P} . We consider the specific situation where the feature θ^* can also be defined as the shared minimizer over Θ of all the risks induced by a loss ℓ and all the conditional marginal laws of $O_{\alpha,t}$ given $Z_{\alpha,t}$ (“all” refers to all $\alpha \in \mathcal{A}$ and $t \geq 1$).

Generally, we make the following assumption (hereafter, for any set S , \mathbb{R}^S denotes the set of functions mapping S to \mathbb{R}).

Assumption 2. There exists a loss function $\ell : \Theta \rightarrow \mathbb{R}^{\mathcal{O} \times \mathcal{Z}}$ that *identifies* the feature of interest θ^* in the sense that θ^* minimizes all the risks $\theta \mapsto \mathbb{E}[\ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}) | Z_{\alpha,t}, F_{t-1}]$ over Θ , “all” referring to all $\alpha \in \mathcal{A}$ and $t \geq 1$. Moreover, for every $\theta \in \Theta$ and sequence $(\theta_t)_{t \geq 1}$ of elements of Θ adapted to $(F_t)_{t \geq 1}$ (*i.e.*, such that each θ_t is F_t -measurable), for all $t \geq 2$ and non-negative integers $\varepsilon_1, \varepsilon_2$ such that $\varepsilon_1 + \varepsilon_2 = 2$,

$$\begin{aligned} & \mathbb{E} \left[\sum_{\alpha \in \mathcal{A}} (\ell(\theta_{t-1})(O_{\alpha,t}, Z_{\alpha,t}))^{\varepsilon_1} \times (\ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}))^{\varepsilon_2} \middle| \bar{Z}_t, F_{t-1} \right] \\ &= \sum_{\alpha \in \mathcal{A}} \mathbb{E} [(\ell(\theta_{t-1})(O_{\alpha,t}, Z_{\alpha,t}))^{\varepsilon_1} \times (\ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}))^{\varepsilon_2} | Z_{\alpha,t}, F_{t-1}]. \end{aligned}$$

Assumption **A2** guarantees some form of stationarity in \mathbb{P} pertaining to its feature of interest θ^* . Thanks to it there is hope that we can learn θ^* from $\bar{O}_1, \dots, \bar{O}_t$ even with t small if $|\mathcal{A}|$ is large (in fact, if the ratio $|\mathcal{A}| / \deg(\mathcal{G})$ is large).

Motivating example of Section 2.3.2. Recall that, in this example, θ^* is the conditional expectation $y \mapsto \theta^*(y|x, z)$ of $Y_{\alpha,t}$ given $(W_{\alpha,t}, X_{\alpha,t}, Z_{\alpha,t}) = (1, x, z)$ (for all (x, z) in the support of any $(X_{\alpha,t}, Z_{\alpha,t})$ conditionally on $W_{\alpha,t} = 1$). Let Θ be the set of measurable functions on $\mathcal{X} \times \mathcal{Z}$ taking their values in $[0, B]$. Thanks to the stationarity assumption made in Section 2.3.2, the least-square loss function $\ell : \Theta \rightarrow \mathbb{R}^{\mathcal{O} \times \mathcal{Z}}$ given by $\ell(\theta) : ((z, x, w, y), z) \mapsto (y - \theta(x, z))^2 \mathbf{1}\{w = 1\}$ (for all $\theta \in \Theta$) identifies θ^* as requested in **A2**. \square

Let $\hat{\theta}_1, \dots, \hat{\theta}_J$ be J algorithms to learn θ^* from $(\bar{O}_t)_{t \geq 1}$. In words, for each $j \in \llbracket J \rrbracket$, $\hat{\theta}_j$ is a procedure that, for every $t \geq 1$, maps $\bar{O}_1, \dots, \bar{O}_t$ to an element of a j -specific subset Θ_j of Θ , namely $\theta_{j,t} \in \Theta_j$ (by convention, $\theta_{j,0}$ is a fixed, pre-specified element of Θ_j). The one-step ahead sequential Super Learner that we are about to introduce is a meta-algorithm that learns, as data accrue, which algorithm in the aforementioned collection performs best.

The measure of performance takes the form of an average cumulative risk conditioned on the sequence $(\bar{Z}_t)_{t \geq 1}$. For every $j \in \llbracket J \rrbracket$, the risk (for short) of $\hat{\theta}_j$ till time $t \geq 1$ is defined as

$$\tilde{R}_{j,t} := \frac{1}{t} \sum_{\tau=1}^t \mathbb{E} \left[\bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) \middle| \bar{Z}_\tau, F_{\tau-1} \right] \quad \text{where} \quad (2.8)$$

$$\bar{\ell}(\theta)(\bar{O}_\tau, \bar{Z}_\tau) := \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \ell(\theta)(O_{\alpha,\tau}, Z_{\alpha,\tau}) \quad \text{for all } \theta \in \Theta, \tau \geq 1. \quad (2.9)$$

The empirical counterpart of (2.8) is

$$\hat{R}_{j,t} := \frac{1}{t} \sum_{\tau=1}^t \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) = \frac{1}{t|\mathcal{A}|} \sum_{\tau=1}^t \sum_{\alpha \in \mathcal{A}} \ell(\theta_{j,\tau-1})(O_{\alpha,\tau}, Z_{\alpha,\tau}). \quad (2.10)$$

Obviously, (2.8) and (2.10) generalize (2.3) and (2.2). At each time $t \geq 1$, the collection of (j, t) -specific empirical risks is minimized at index \hat{j}_t defined in (2.1). The one-step ahead sequential Super Learner is the meta-algorithm that learns θ^* by mapping $\bar{O}_1, \dots, \bar{O}_t$ to $\theta_{\hat{j}_t,t}$ for every $t \geq 1$. To assess how well the one-step ahead sequential Super Learner performs, we compare its risk to that of the oracular algorithm that learns θ^* by mapping $\bar{O}_1, \dots, \bar{O}_t$ to $\theta_{\tilde{j}_t,t}^*$ at each time $t \geq 1$, \tilde{j}_t being defined in (2.4).

Comparing the one-step ahead sequential Super Learner to its oracular counterpart So far we have defined the risks of $\hat{\theta}_1, \dots, \hat{\theta}_J$, see (2.8). By analogy, for every $\theta \in \Theta$ and $t \geq 1$, let the risk of θ at time t be

$$\tilde{R}_t(\theta) := \frac{1}{t} \sum_{\tau=1}^t \mathbb{E} \left[\bar{\ell}(\theta)(\bar{O}_\tau, \bar{Z}_\tau) \middle| \bar{Z}_\tau, F_{\tau-1} \right].$$

The risk $\tilde{R}_t(\theta)$ can be interpreted as the risk till time $t \geq 1$ of a dummy algorithm that constantly maps $\bar{O}_1, \dots, \bar{O}_t$ to θ (the algorithm is said dummy because it does not learn). Let $\theta^\circ \in \Theta$ be such that

$$\tilde{R}_t^\circ := \tilde{R}_t(\theta^\circ) \leq \min_{j \in \llbracket J \rrbracket} \min_{\theta \in \Theta_j} \tilde{R}_t(\theta).$$

Under **A2**, θ° could be set to θ^* , but other choices might be made on a case by case basis. Our main results compare the excess risks of the one-step ahead sequential Super Learner and of the oracle, that is, they compare

$$\tilde{R}_{\hat{j}_t,t} - \tilde{R}_t^\circ \geq 0 \quad \text{to} \quad \tilde{R}_{\tilde{j}_t,t} - \tilde{R}_t^\circ \geq 0.$$

They rely on the following assumptions.

For every $\theta \in \Theta$, let $\Delta^\circ \ell(\theta) := \ell(\theta) - \ell(\theta^\circ)$.

Assumption 3. There exists $b_1 > 0$ such that $\sup_{\theta \in \Theta} \|\Delta^\circ \ell(\theta)\|_\infty \leq b_1$. Moreover there exists $b_2 \in]0, 2b_1]$ such that, almost surely, for all $\alpha \in \mathcal{A}$, $t \geq 1$ and $\theta \in \Theta$,

$$|\Delta^\circ \ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}) - \mathbb{E}[\Delta^\circ \ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}) | Z_{\alpha,t}, F_{t-1}]| \leq b_2.$$

Assumption 4. There exist $\beta \in]0, 1]$ and $\gamma > 0$ such that, almost surely, for all $\alpha \in \mathcal{A}$, $t \geq 1$ and $\theta \in \Theta$,

$$\mathbb{E} \left[\left(\Delta^\circ \ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}) \right)^2 \middle| Z_{\alpha,t}, F_{t-1} \right] \leq \gamma \left(\mathbb{E}[\Delta^\circ \ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}) | Z_{\alpha,t}, F_{t-1}] \right)^\beta.$$

Assumption 5. There exists $v_1 > 0$ such that, almost surely, for all $\alpha \in \mathcal{A}$, $t \geq 1$ and $\theta \in \Theta$,

$$\text{Var}[\Delta^\circ \ell(\theta)(O_{\alpha,t}, Z_{\alpha,t}) | Z_{\alpha,t}, F_{t-1}] \leq v_1.$$

Assumption **A4** is a so-called ‘‘variance bound’’, a well-known concept in statistical learning theory [Bartlett et al., 2005, Koltchinskii, 2006, Bartlett et al., 2006]. Under **A3**, the radius of the loss class is bounded. Note that **A3** implies **A5** with an upper-bound equal to $(b_2)^2$, which will typically be much larger than v_1 .

Motivating example of Section 2.3.2. In the example, we can choose $\theta^\circ := \theta^*$ and **A3**, **A4** (with $\beta = 1$) and **A5** are met. The fact that **A4** is met with $\beta = 1$ follows from the classical argument of strong convexity recalled for self-containedness in Appendix 2.8. \square

Fix arbitrarily $a, x > 0$ and $t \geq 1$. Stated in Theorem 1, the master result consists of two upper bounds on

$$\mathbb{P} \left[\tilde{R}_{\hat{j}_t, t} - \tilde{R}_t^\circ \geq (1 + 2a) \left(\tilde{R}_{\hat{j}_t, t} - \tilde{R}_t^\circ \right) + x \right]. \quad (2.11)$$

They yield two upper bounds on

$$\mathbb{E} \left[\tilde{R}_{\hat{j}_t, t} - \tilde{R}_t^\circ - (1 + 2a) \left(\tilde{R}_{\hat{j}_t, t} - \tilde{R}_t^\circ \right) \right] \quad (2.12)$$

which are stated in Corollary 2.

Theorem 1 (High probability oracular inequalities). *Suppose that **A1**, **A2**, **A3**, **A4** and **A5** are met. There exist $C_1(a)$, $C_2(a)$, $C'_1(a)$, $C'_2(a) > 0$ such that, for any two integers $N, N' \geq 2$, if $x \geq \underline{x}(a, N)$ then (2.11) is smaller than*

$$2JN \left[\exp \left(-\frac{tx^{2-\beta}}{C_1(a)} \right) + \exp \left(-\frac{tx}{C_2(a)} \right) \right] \quad (2.13)$$

and if $x \geq \underline{x}'(a, N')$, then (2.11) is smaller than

$$2e^2 JN' \left[\exp \left(-\frac{[|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))]x^{2-\beta}}{C'_1(a)} \right) + \exp \left(-\frac{[|\mathcal{A}|/\deg(\mathcal{G})]x}{C'_2(a)} \right) \right]. \quad (2.14)$$

Corollary 2 (Oracular inequalities for the expected excess risk). *Suppose that **A1**, **A2**, **A3**, **A4** and **A5** are met. There exists $C_3, C'_3 > 0$ such that, for any $a \in]0, 1]$, if $N \geq 2$ satisfies*

$$N \geq \frac{\beta}{2 - \beta} \frac{\log(t) + \log(C_3)}{\log(2)}, \quad (2.15)$$

then (2.12) is smaller than

$$3 \left(\frac{C_1(a)}{t} \log(2JN) \right)^{1/(2-\beta)} + \frac{2C_2(a)}{t} \log(2JN) \quad (2.16)$$

and if $N \geq 2$ satisfies

$$N' \geq \frac{\beta}{2 - \beta} \frac{\log(|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))) + \log(C'_3)}{\log(2)}, \quad (2.17)$$

then (2.12) is smaller than

$$3 \left(\frac{C'_1(a)}{|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))} \log(2JN') \right)^{1/(2-\beta)} + \frac{2C'_2(a)}{|\mathcal{A}|/\deg(\mathcal{G})} \log(2JN'). \quad (2.18)$$

Theorem 1 and Corollary 2 generalize the results of Benkeser et al. [2018] in two aspects. First, they do not require assumptions akin to *their assumptions* A3 and A4, which are meant to deal with the randomness at play in $\tilde{R}_{j,t}$ and in $\mathbb{V}\text{ar}[\Delta^\circ \ell(O_{\alpha,t}, Z_t) | Z_t, F_{t-1}]$. Instead we exploit a so-called stratification argument inspired by Cesa-Bianchi and Gentile [2008]. Second, our results leverage the use of a conditional dependency graph to model the amount of conditional independence within the data-structures $\bar{O}_1, \dots, \bar{O}_t$. Before giving a few more details on the proofs, let us compare the bounds (2.13) and (2.14), (2.16) and (2.18).

Leveraging a large ratio $|\mathcal{A}|/\deg(\mathcal{G})$ in the face of a small t The real numbers $C_1(a), C_2(a), C'_1(a), C'_2(a), C_3, C'_3, \underline{x}(a, N), \underline{x}'(a, N')$ and an additional $v_2 > 0$ depend on the constants introduced in **A1**, **A3**, **A4**, **A5**. Their definitions are given in Section 2.7 – specifically, at the end of Step 2 (v1) and Step 2 (v2) in Section 2.7.1, at the end of the proof of Corollary 2 in Section 2.7.2, and in (2.23) for v_2 .

If one chooses $N = N'$ in (2.16) and (2.18), then it is easy to check that the two terms in the RHS expression of (2.18) are smaller than their counterparts in (2.16) if and only if

$$t^{1+\beta} \leq \frac{|\mathcal{A}|/\deg(\mathcal{G})}{2e^{28\beta}} \quad \text{and} \quad t \leq \frac{|\mathcal{A}|/\deg(\mathcal{G})}{45e/2}. \quad (2.19)$$

Furthermore, a simple sufficient condition for (2.19) to be met is

$$t^{1+\beta} \leq \frac{|\mathcal{A}|/\deg(\mathcal{G})}{2e^{28\beta}} \quad \text{and} \quad \frac{|\mathcal{A}|}{\deg(\mathcal{G})} \geq 24e \times (3/(2e))^{1/\beta}. \quad (2.20)$$

Thus, if (2.20) is met (note that $24e \times 3/(2e) = 36 \geq 24e \times (3/(2e))^{1/\beta}$ whatever is $\beta \in]0, 1[$) and if we make the following (valid) choices in Corollary 2,

$$N = N' \geq \frac{\beta}{2 - \beta} \frac{\log(|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))) + \log(C'_3) + \left(\log[C_3/(2e^{28^\beta} C'_3)]\right)_+,}{\log(2)},$$

then the oracular inequalities (2.16) and (2.18) for the expected risk hold true, the latter being sharper than the former. In words, our analysis does take advantage of the fact that $|\mathcal{A}|/\deg(\mathcal{G})$ is large in the face of t being comparatively small.

A few details on the proofs Theorem 1 notably hinges on the Fan-Grama-Liu concentration inequality for martingales [Theorem 3.10 in Bercu et al., 2015] and on the following result, tailored to our needs and derived from a concentration inequality for sums of partly dependent random variables shown by Janson [2004]. For each $j \in \llbracket J \rrbracket$ and $t \geq 1$, introduce the two (j, t) -specific averages of conditional variances

$$\text{var}_{j,t} := \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \text{Var} [\Delta^\circ \ell(\theta_{j,\tau-1})(O_{\alpha,t} Z_{\alpha,t}) | Z_{\alpha,t}, F_{t-1}], \quad (2.21)$$

$$\widetilde{\text{var}}_{j,t} := \frac{1}{t} \sum_{\tau=1}^t \text{Var} [\Delta^\circ \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) | \bar{Z}_\tau, F_{\tau-1}]. \quad (2.22)$$

Define

$$v_2 := \frac{3\pi}{2} \left[\left(\frac{15b_2}{|\mathcal{A}|/\deg(\mathcal{G})} \right)^2 + \frac{64v_1}{|\mathcal{A}|/\deg(\mathcal{G})} \right]. \quad (2.23)$$

Theorem 3. *Suppose that **A3** and **A4** are met. For each $j \in \llbracket J \rrbracket$, $\widetilde{\text{var}}_{j,t} \leq v_2$ almost surely. Moreover, for any $V > 0$ and all $x \geq 0$, if*

$$\widetilde{\mathcal{F}}_V := \left[\max_{\tau \in \llbracket t \rrbracket} \{\text{var}_{j,\tau}\} \leq V \right], \quad (2.24)$$

then

$$\mathbb{P} \left[|[\widehat{R}_{j,t} - \widehat{R}_t^\circ] - [\widetilde{R}_{j,t} - \widetilde{R}_t^\circ]| \geq x, \widetilde{\mathcal{F}}_V \right] \leq \exp \left(2 - \frac{[|\mathcal{A}|/\deg(\mathcal{G})]x^2}{32e^2V + 15eb_2x} \right). \quad (2.25)$$

Our proof of Theorem 3 consists in deriving a Rosenthal inequality from Janson's concentration inequality [2004], following Petrov's line of proof [1995], in using a convexity argument, then in applying the same method as in [Dedecker, 2001, Corollary 3(b)] (inspired by the proof of Theorem 6 in [Doukhan et al., 1984]). Inequality (2.25) plays a key role in the derivation of (2.18). The fact that the first term in the RHS expression in (2.18) features $|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))$ and not $|\mathcal{A}|/\deg(\mathcal{G})$ may be deemed pessimistic but is inherent to our scheme of proof. Note that substituting a sharp Marcinkiewicz-Zygmund-like inequality [Rio, 2009, Theorem 2.9] for the convexity argument that leads to (2.53) does not solve the issue.

Furthermore, it is noteworthy that our results extend seamlessly to the case that every expression $\ell(\theta_{\tau-1})(O_{\alpha,\tau}, Z_{\alpha,\tau})$ with $\theta_{\tau-1}$ $F_{\tau-1}$ -measurable is replaced by an expression of the form $\ell(\theta_{\tau-1})(O_{\alpha,\tau}, Z_{\alpha,\tau}) \times \omega_\tau(O_{\alpha,\tau}, Z_{\alpha,\tau})$, where ω_τ is a $F_{\tau-1}$ -measurable weighting function. This proves very useful in the context of reinforcement learning, allowing to rely on importance sampling weighting.

2.7 Appendix: proofs

2.7.1 Proof of Theorem 1

The proof unfolds in three steps.

Step 1: an algebraic decomposition For all $j \in \llbracket J \rrbracket$, $t \geq 1$ and $\theta \in \Theta$, let us define

$$\tilde{H}_{j,t} := \tilde{R}_{j,t} - \tilde{R}_t^\circ, \quad \hat{H}_{j,t} := \hat{R}_{j,t} - \hat{R}_t(\theta^\circ) \quad \text{and} \quad \Delta^\circ \bar{\ell}(\theta)(\bar{O}_t, \bar{Z}_t) := \bar{\ell}(\theta)(\bar{O}_t, \bar{Z}_t) - \bar{\ell}(\theta^\circ)(\bar{O}_t, \bar{Z}_t)$$

($\bar{\ell}(\theta)$ is defined in (2.9)). Fix arbitrarily $a > 0$. An algebraic decomposition at the heart of all studies of the Super Learner [see, *e.g.*, Dudoit and van der Laan, 2005, van der Laan et al., 2007, Benkeser et al., 2018]) states that the excess risk of the Super Learner (that is, $\tilde{H}_{j,t,t}$) can be bounded by $(1 + 2a)$ times the excess risk of the oracle (that is, $\tilde{H}_{j,t,t}^\circ$), plus some remainder terms:

$$\begin{aligned} \tilde{H}_{j,t,t} &\leq (1 + 2a) \tilde{H}_{j,t,t}^\circ + A_{j,t,t}^\circ(a) + B_{j,t,t}^\circ(a) \\ &\leq (1 + 2a) \tilde{H}_{j,t,t}^\circ + \max_{j \in \llbracket J \rrbracket} \{A_{j,t,t}^\circ(a)\} + \max_{j \in \llbracket J \rrbracket} \{B_{j,t,t}^\circ(a)\} \end{aligned} \quad (2.26)$$

where

$$A_{j,t,t}(a) := (1 + a) \left(\tilde{H}_{j,t,t} - \hat{H}_{j,t,t} \right) - a \tilde{H}_{j,t,t} \quad \text{and} \quad B_{j,t,t}(a) := (1 + a) \left(\hat{H}_{j,t,t} - \tilde{H}_{j,t,t} \right) - a \tilde{H}_{j,t,t}.$$

The first terms in the definitions of $A_{j,t,t}(a)$ and $B_{j,t,t}(a)$ equal $\pm(1 + a)$ times

$$\frac{1}{t} \sum_{\tau=1}^t \left(\Delta^\circ \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) - \mathbb{E} \left[\Delta^\circ \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) \mid \bar{Z}_\tau, F_{\tau-1} \right] \right),$$

that is as the average of the t first terms of a martingale difference sequence. As for the shared second term in the definitions of $A_{j,t,t}(a)$ and $B_{j,t,t}(a)$, it satisfies $-a \tilde{H}_{j,t,t} \leq 0$. The second step of the proof consists in exploiting two so-called Bernstein's inequalities to control the probabilities $\mathbb{P}[A_{j,t,t}(a) \geq x]$ and $\mathbb{P}[B_{j,t,t}(a) \geq x]$ for $x \geq 0$.

Step 2: Bounding positive deviations of $A_{j,t,t}(a)$ and $B_{j,t,t}(a)$ Set arbitrarily two integers $N, N' \geq 2$ and a real number $x \geq 0$. The analysis of $\mathbb{P}[B_{j,t,t}(a) \geq x]$ is exactly the same as that of $\mathbb{P}[A_{j,t,t}(a) \geq x]$, so we present only the latter. The key to the analysis is a so-called stratification argument inspired by Cesa-Bianchi and Gentile [2008].

For every $j \in \llbracket J \rrbracket$ and $t \geq 1$, recall the definitions (2.21) and (2.22) of $\text{var}_{j,t}$ and $\widetilde{\text{var}}_{j,t}$. On the one hand, by **A4** and because the functions of a real variable $u \mapsto u^2$ and $u \mapsto u^\beta$ are respectively convex and concave, it holds that

$$\begin{aligned} \widetilde{\text{var}}_{j,t} &\leq \frac{1}{t} \sum_{\tau=1}^t \mathbb{E} \left[\left(\Delta^\circ \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) \right)^2 \middle| \bar{Z}_\tau, F_{\tau-1} \right] \\ &\leq \gamma \left(\frac{1}{t} \sum_{\tau=1}^t \mathbb{E} \left[\Delta^\circ \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) \middle| \bar{Z}_\tau, F_{\tau-1} \right] \right)^\beta = \gamma \left(\tilde{H}_{j,t} \right)^\beta \end{aligned} \quad (2.27)$$

almost surely. Moreover, it also holds that $\widetilde{\text{var}}_{j,t} \leq v_2$ almost surely by Theorem 3. The previous upper bound and (2.27) play a key role in the first version of Step 2 (Step 2 (v1)) presented below. On the other hand, by **A2**, **A4**, and because the function $u \mapsto u^\beta$ is concave it holds almost surely that, for all $\tau \in \llbracket t \rrbracket$,

$$\begin{aligned} \text{var}_{j,\tau} &\leq \frac{1}{|\mathcal{A}|} \sum_{\alpha \in \mathcal{A}} \mathbb{E} \left[\left(\Delta^\circ \ell(\theta_{j,\tau-1})(O_{\alpha,\tau}, Z_{\alpha,\tau}) \right)^2 \middle| Z_{\alpha,\tau}, F_{\tau-1} \right] \\ &\leq \gamma \left(\mathbb{E} \left[\Delta^\circ \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) \middle| \bar{Z}_\tau, F_{\tau-1} \right] \right)^\beta. \end{aligned}$$

Consequently if $\tilde{H}_{j,t} \leq B$ (an inequality that holds almost surely when $B = b_1$, by **A3**), then it also holds that

$$B \geq \tilde{H}_{j,t} = \frac{1}{t} \sum_{\tau=1}^t \mathbb{E} \left[\Delta^\circ \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_\tau, \bar{Z}_\tau) \middle| \bar{Z}_\tau, F_{\tau-1} \right] \geq \frac{1}{t} \sum_{\tau=1}^t (\text{var}_{j,\tau} / \gamma)^{1/\beta}.$$

In summary we will use that, for any $B > 0$,

$$\mathbf{1} \left\{ \tilde{H}_{j,t} \leq B \right\} \leq \mathbf{1} \left\{ \max_{1 \leq \tau \leq t} \{\text{var}_{j,\tau}\} \leq \gamma (tB)^\beta \right\} = \mathbf{1} \left\{ \tilde{\mathcal{F}}_{\gamma(tB)^\beta} \right\} \quad (2.28)$$

($\tilde{\mathcal{F}}_V$ is defined for any $V > 0$ in (2.24)). The upper bound $\tilde{H}_{j,t} \leq b_1$ and (2.28) play a key role in the second version of Step 2 (Step 2 (v2)) presented below.

Step 2 (v1). Set $v_2^{(-1)} := 0$ and, for all $i \in \llbracket N-1 \rrbracket$, $v_2^{(i)} := 2^{i+1-N} \times v_2$. In view of (2.27) and since $\widetilde{\text{var}}_{j,t} \in \cup_{i=0}^N [v_2^{(i-1)}, v_2^{(i)}]$ almost surely, it holds that

$$\begin{aligned} \mathbb{P} [A_{j,t}(a) \geq x] &= \mathbb{P} \left[\tilde{H}_{j,t} - \hat{H}_{j,t} \geq \frac{1}{1+a} \left(x + a \tilde{H}_{j,t} \right) \right] \\ &\leq \mathbb{P} \left[\tilde{H}_{j,t} - \hat{H}_{j,t} \geq \frac{1}{1+a} \left(x + a (\widetilde{\text{var}}_{j,t} / \gamma)^{1/\beta} \right) \right] \\ &\leq \sum_{i=0}^{N-1} \mathbb{P} \left[\tilde{H}_{j,t} - \hat{H}_{j,t} \geq \frac{1}{1+a} \left(x + a (\widetilde{\text{var}}_{j,t} / \gamma)^{1/\beta} \right), \right. \\ &\quad \left. \widetilde{\text{var}}_{j,t} \in [v_2^{(i-1)}, v_2^{(i)}] \right] \end{aligned}$$

$$\leq \sum_{i=0}^{N-1} \mathbb{P} \left[\widetilde{H}_{j,t} - \widehat{H}_{j,t} \geq \frac{1}{1+a} \left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta} \right), \right. \\ \left. \widetilde{\text{var}}_{j,t} \leq v_2^{(i)} \right]. \quad (2.29)$$

Note that $(\widetilde{H}_{j,t} - \widehat{H}_{j,t})_{t \geq 1}$ is a martingale adapted to the filtration $(\sigma(F_t, \sigma(\bar{Z}_{t+1})))_{t \geq 1}$. By **A3** and the Fan-Grama-Liu concentration inequality for martingales [Theorem 3.10 in Bercu et al., 2015], (2.29) implies

$$\mathbb{P}[A_{j,t}(a) \geq x] \leq \sum_{i=0}^{N-1} \exp \left(-\frac{1}{2} \frac{tD_i(x)}{(1+a)^2} \right), \quad (2.30)$$

where, for all $i \in \llbracket N-1 \rrbracket$,

$$D_i(x) := \frac{\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta} \right)^2}{v_2^{(i)} + \frac{1}{3} \frac{b_2}{1+a} \left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta} \right)}.$$

Set arbitrarily $i \in \llbracket N-1 \rrbracket \cup \{0\}$ and define $x_i := 3(1+a)v_2^{(i)}/b_2 - a(v_2^{(i-1)}/\gamma)^{1/\beta}$.

— If $x \leq x_i$, then $v_2^{(i)} \geq (x + a(v_2^{(i-1)}/\gamma)^{1/\beta}) \times b_2/(3(1+a))$ hence

$$D_i(x) \geq \frac{\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta} \right)^2}{2v_2^{(i)}} = \frac{\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta} \right)^{2-\beta}}{2v_2^{(i)} / \left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta} \right)^\beta} \\ \geq \frac{x^{2-\beta}}{2v_2^{(i)} / \left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta} \right)^\beta}. \quad (2.31)$$

If $i \neq 0$, then (2.31) entails

$$D_i(x) \geq \frac{x^{2-\beta}}{2\gamma v_2^{(i)} / (a^\beta v_2^{(i-1)})} = \frac{x^{2-\beta}}{4\gamma/a^\beta}. \quad (2.32)$$

If $i = 0$, then (2.32) is also met if and only if $x \geq \underline{x}(a, N)$, where $\underline{x}(a, N) := a[2^{-N}v_2/\gamma]^{1/\beta}$.

— Moreover if $x \geq x_i$, then $v_2^{(i)} \leq (x + a(v_2^{(i-1)}/\gamma)^{1/\beta}) \times b_2/(3(1+a))$ hence

$$D_i(x) \geq \frac{\left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta} \right)^2}{\frac{2}{3} \frac{b_2}{1+a} \left(x + a(v_2^{(i-1)}/\gamma)^{1/\beta} \right)} = \frac{x + a(v_2^{(i-1)}/\gamma)^{1/\beta}}{\frac{2}{3} \frac{b_2}{1+a}} \geq \frac{x}{\frac{2}{3} \frac{b_2}{1+a}}. \quad (2.33)$$

Therefore, in light of (2.30), (2.32), (2.33), if $C_1(a) := 2^{5-\beta}(1+a)^2\gamma/a^\beta$ and $C_2(a) := 8(1+a)b_2/3$, then it holds for all $x \geq \underline{x}(a, N)$ that

$$\mathbb{P}[A_{j,t}(a) \geq x] \leq \sum_{i=0}^{N-1} N-1 \left[\mathbf{1}\{x \leq x_i\} \exp \left(-\frac{t \times (2x)^{2-\beta}}{C_1(a)} \right) \right]$$

$$\begin{aligned}
& + \mathbf{1}\{x \geq x_i\} \exp\left(-\frac{t \times (2x)}{C_2(a)}\right) \Big] \\
& \leq N \left[\exp\left(-\frac{t \times (2x)^{2-\beta}}{C_1(a)}\right) + \exp\left(-\frac{t \times (2x)}{C_2(a)}\right) \right]. \quad (2.34)
\end{aligned}$$

Step 2 (v2). This step is very similar to Step 2 (v1). Set $b_1^{(-1)} := 0$ and, for all $i \in \llbracket N'-1 \rrbracket$, $b_1^{(i)} := 2^{i+1-N'} \times b_1$. In view of (2.28) and since $\tilde{H}_{j,t} \in \cup_{i=0}^{N'-1} [b_1^{(i-1)}, b_1^{(i)}]$ almost surely, it holds that

$$\begin{aligned}
\mathbb{P}[A_{j,t}(a) \geq x] &= \mathbb{P}\left[\tilde{H}_{j,t} - \hat{H}_{j,t} \geq \frac{1}{1+a} (x + a\tilde{H}_{j,t})\right] \\
&\leq \sum_{i=0}^{N'-1} \mathbb{P}\left[\tilde{H}_{j,t} - \hat{H}_{j,t} \geq \frac{1}{1+a} (x + a\tilde{H}_{j,t}), \right. \\
&\quad \left. \tilde{H}_{j,t} \in [b_1^{(i-1)}, b_1^{(i)}]\right] \\
&\leq \sum_{i=0}^{N'-1} \mathbb{P}\left[\tilde{H}_{j,t} - \hat{H}_{j,t} \geq \frac{1}{1+a} (x + ab_1^{(i-1)}), \tilde{H}_{j,t} \leq b_1^{(i)}\right] \\
&\leq \sum_{i=0}^{N'-1} \mathbb{P}\left[\tilde{H}_{j,t} - \hat{H}_{j,t} \geq \frac{1}{1+a} (x + ab_1^{(i-1)}), \tilde{\mathcal{F}}_{\gamma(tb_1^{(i)})^\beta}\right]. \quad (2.35)
\end{aligned}$$

By **A3** and **A4**, Theorem 3 applies and (2.35) yields

$$\mathbb{P}[A_{j,t}(a) \geq x] \leq \sum_{i=0}^{N'-1} \exp\left(2 - \frac{|\mathcal{A}|/\deg(\mathcal{G})}{(1+a)^2} D'_i(x)\right), \quad (2.36)$$

where, for all $i \in \llbracket N'-1 \rrbracket$,

$$D'_i(x) := \frac{(x + ab_1^{(i-1)})^2}{32e^2\gamma(tb_1^{(i)})^\beta + \frac{15eb_2}{1+a} (x + ab_1^{(i-1)})}.$$

Set arbitrarily $i \in \llbracket N'-1 \rrbracket \cup \{0\}$ and define $x'_i := 32e(1+a)\gamma(tb_1^{(i)})^\beta / (15b_2) - ab_1^{(i-1)}$.

— If $x \leq x'_i$, then $32e^2\gamma(tb_1^{(i)})^\beta \geq (x + ab_1^{(i-1)}) \times 15eb_2 / (1+a)$ hence

$$\begin{aligned}
D'_i(x) &\geq \frac{(x + ab_1^{(i-1)})^2}{64e^2\gamma(tb_1^{(i)})^\beta} = \frac{(x + ab_1^{(i-1)})^{2-\beta}}{64e^2\gamma(tb_1^{(i)})^\beta / (x + ab_1^{(i-1)})^\beta} \\
&\geq \frac{x^{2-\beta}}{64e^2\gamma(tb_1^{(i)})^\beta / (x + ab_1^{(i-1)})^\beta}. \quad (2.37)
\end{aligned}$$

If $i \neq 0$, then (2.37) entails

$$D'_i(x) \geq \frac{x^{2-\beta}}{64e^2\gamma(tb_1^{(i)})^\beta/(ab_1^{(i-1)})^\beta} = \frac{x^{2-\beta}}{64e^2\gamma(2t/a)^\beta}. \quad (2.38)$$

If $i = 0$, then (2.38) is also met if and only if $x \geq \underline{x}'(a, N') := ab_1 2^{-N'}$.

— Moreover if $x \geq x'_i$, then $32e^2\gamma(tb_1^{(i)})^\beta \leq (x + ab_1^{(i-1)}) \times 15eb_2/(1+a)$ hence

$$D'_i(x) \geq \frac{(x + ab_1^{(i-1)})^2}{\frac{30eb_2}{1+a} (x + ab_1^{(i-1)})} = \frac{x + ab_1^{(i-1)}}{\frac{30eb_2}{1+a}} \geq \frac{x}{\frac{30eb_2}{1+a}}. \quad (2.39)$$

Therefore, in light of (2.36), (2.38), (2.39), if $C'_1(a) := 2^{6+2\beta}e^2(1+a)^2\gamma/a^\beta$ and $C'_2(a) := 60e(1+a)b_2$, then it holds for all $x \geq \underline{x}'(a, N')$ that

$$\begin{aligned} \mathbb{P}[A_{j,t}(a) \geq x] &\leq \sum_{i=0}^{N'-1} \left[\mathbf{1}\{x \leq x'_i\} \exp\left(2 - \frac{[|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))] \times (2x)^{2-\beta}}{C'_1(a)}\right) \right. \\ &\quad \left. + \mathbf{1}\{x \geq x'_i\} \exp\left(2 - \frac{[|\mathcal{A}|/\deg(\mathcal{G})] \times (2x)}{C'_2(a)}\right) \right] \\ &\leq N' \left[\exp\left(2 - \frac{[|\mathcal{A}|/(t^\beta \deg(\mathcal{G}))] \times (2x)^{2-\beta}}{C'_1(a)}\right) \right. \\ &\quad \left. + \exp\left(2 - \frac{[|\mathcal{A}|/\deg(\mathcal{G})] \times (2x)}{C'_2(a)}\right) \right]. \end{aligned} \quad (2.40)$$

Step 3: end of the proof In view of (2.26), a union bound implies that

$$\begin{aligned} \mathbb{P}\left[\tilde{H}_{j_t,t} - (1+2a)\tilde{H}_{j_t,t} \geq x\right] &\leq \mathbb{P}\left[\max_{j \in [J]} \{A_{j,t}(a)\} + \max_{j \in [J]} \{B_{j,t}(a)\} \geq x\right] \\ &\leq \sum_{j=1}^J (\mathbb{P}[A_{j,t}(a) \geq x/2] + \mathbb{P}[B_{j,t}(a) \geq x/2]). \end{aligned}$$

For all $x \geq \underline{x}(a, N)$, (2.13) follows from (2.34) and the above inequality; for all $x \geq \underline{x}'(a, N')$, (2.14) follows from (2.40) and the above inequality. This completes the proof of Theorem 1. \square

2.7.2 Proof of Corollary 2

Corollary 2 follows from the straightforward application, twice, of the next technical lemma, based on (2.13) on the one hand and on (2.14) on the other hand.

Lemma 4. Let $a, b, c > 0$, $\beta \in]0, 1]$ be some constants and $(\underline{x}(N))_{N \geq 2}$ be a sequence of positive numbers that decreases to 0. Let U be a real valued random variable such that $E[|U|] < \infty$ and, for all integer $N \geq 2$ and all $x \geq \underline{x}(N) > 0$,

$$\mathbb{P}[U \geq x] \leq aN \left[\exp(-x^{2-\beta}/b) + \exp(-x/c) \right]. \quad (2.41)$$

If $N \geq \min\{n \geq 2 : \underline{x}(n) \leq b^{1/(2-\beta)}, \log(an) \geq 1\}$, then

$$\mathbb{E}[U] \leq 3(b \log(aN))^{1/(2-\beta)} + 2c \log(aN). \quad (2.42)$$

Proof of Lemma 4. It is well known that

$$\mathbb{E}[U] \leq \mathbb{E}[U \mathbf{1}\{U \geq 0\}] = \int_0^\infty \mathbb{P}[U \mathbf{1}\{U \geq 0\} \geq x] dx = \int_0^\infty \mathbb{P}[U \geq x] dx$$

and that $\min\{1, a+b\} \leq \min\{1, a\} + \min\{1, b\}$. Therefore, for any $N \geq 2$,

$$\begin{aligned} \mathbb{E}[U] &\leq \int_0^\infty \left(\mathbf{1}\{x < \underline{x}(N)\} + \mathbf{1}\{x \geq \underline{x}(N)\} \min\left\{1, \right. \right. \\ &\quad \left. \left. aN \left[\exp(-x^{2-\beta}/b) + \exp(-x/c) \right] \right\} \right) dx \\ &\leq \underline{x}(N) + \int_0^\infty \min\{1, aN \exp(-x^{2-\beta}/b)\} dx + \int_0^\infty \min\{1, aN \exp(-x/c)\} dx. \end{aligned} \quad (2.43)$$

Let \mathcal{L} and \mathcal{R} be the above LHS and RHS integrals. Choose $N \geq \min\{n \geq 2 : \underline{x}(n) \leq b^{1/(2-\beta)}, \log(an) \geq 1\}$.

Bounding \mathcal{L} . Let $x_{\mathcal{L}}$ be chosen so that

$$aN \exp(-x_{\mathcal{L}}^{2-\beta}/b) = 1, \quad \text{i.e.,} \quad x_{\mathcal{L}} := (b \log(aN))^{1/(2-\beta)}.$$

Now, thanks to the change of variable $u = x^{2-\beta}/b$ and because $u \mapsto u^{1/(2-\beta)-1}$ is nonincreasing,

$$\begin{aligned} \mathcal{L} &= x_{\mathcal{L}} + aN \int_{x_{\mathcal{L}}}^\infty \exp(-x^{2-\beta}/b) dx \\ &= x_{\mathcal{L}} + aN b^{1/(2-\beta)} \int_{\log(aN)}^\infty \exp(-u) u^{1/(2-\beta)-1} du \\ &\leq x_{\mathcal{L}} + \frac{aN (b \log(aN))^{1/(2-\beta)}}{\log(aN)} \int_0^\infty \exp(-u) du \\ &= x_{\mathcal{L}} (1 + 1/\log(aN)) \leq 2(b \log(aN))^{1/(2-\beta)}. \end{aligned} \quad (2.44)$$

Bounding \mathcal{R} . Let $x_{\mathcal{R}}$ be chosen so that $aN \exp(-x_{\mathcal{R}}/c) = 1$, i.e., $x_{\mathcal{R}} := c \log(aN)$. It is readily seen that

$$\mathcal{R} = x_{\mathcal{R}} + aN \int_{x_{\mathcal{R}}}^\infty \exp(-x/c) dx = x_{\mathcal{R}} + acN \exp(-x_{\mathcal{R}}/c) = c(1 + \log(aN)). \quad (2.45)$$

In view of (2.43), (2.44), (2.45), and by choice of N , we obtain

$$\begin{aligned}\mathbb{E}[U] &\leq b^{1/(2-\beta)} + 2(b \log(aN))^{1/(2-\beta)} + c(1 + \log(aN)) \\ &\leq 3(b \log(aN))^{1/(2-\beta)} + 2c \log(aN).\end{aligned}$$

This completes the proof. \square

Set $t \geq 1$ and $a \in]0, 1]$. In view of (2.13), Lemma 4 yields (2.16) under the sufficient condition that $N \geq 2$ also satisfy (2.15) with $C_3 := (v_2/\gamma)^{(2-\beta)/\beta}/(2^{5-\beta}\gamma)$. Moreover, in view of (2.14), Lemma 4 also yields (2.18) under the sufficient condition that $N \geq 2$ also satisfy (2.17) with $C'_3 := b_1/(2^{6+2\beta}e^2\gamma)$. This completes the proof of the corollary. \square

2.7.3 Proof of Theorem 3

The proof of Theorem 3 hinges on a Bernstein-like concentration inequality for sums of partly dependent random variables shown by Janson [2004, Theorem 2.3]. Janson emphasizes that his theorem uses the independence of suitable (large) subsets of $(\zeta_\alpha)_{\alpha \in \mathcal{A}}$, not any other information on the dependencies, so that the result must be suboptimal when the dependencies that exist are weak. We recall the theorem for completeness.

Theorem 5 (Janson [2004]). *Let $(\zeta_\alpha)_{\alpha \in \mathcal{A}}$ be a collection of random variables with dependency graph \mathcal{G} such that $\zeta_\alpha - \mathbb{E}[\zeta_\alpha] \leq B$ for some $B > 0$ and all $\alpha \in \mathcal{A}$. Define $\mathcal{Z} := |\mathcal{A}|^{-1} \sum_{\alpha \in \mathcal{A}} \zeta_\alpha$ and $V := |\mathcal{A}|^{-1} \sum_{\alpha \in \mathcal{A}} \text{Var}[\zeta_\alpha]$. Then, for all $x \geq 0$,*

$$\mathbb{P}[\mathcal{Z} - \mathbb{E}[\mathcal{Z}] \geq x] \leq \exp\left(-\frac{|\mathcal{A}|V}{B^2 \deg(\mathcal{G})} h\left(\frac{4Bx}{5V}\right)\right), \quad (2.46)$$

where $h : u \mapsto (1+u) \log(1+u) - u$.

Note that (2.25) from Theorem 3 also writes as

$$\mathbb{P}\left[|\hat{H}_{j,t} - \tilde{H}_{j,t}| \geq x, \tilde{\mathcal{F}}_V\right] \leq \exp\left(2 - \frac{[|\mathcal{A}|/\deg(\mathcal{G})]x^2}{32e^2V + 15eb_2x}\right).$$

Following the line of proof of the Rosenthal inequality by Petrov [1995, page 59] (see also the proof of Theorem 5.2 in [Baraud, 2000]), we use (2.46) to control $\mathbb{E}[|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]|^p]$ hence $\mathbb{E}[|\hat{H}_{j,t} - \tilde{H}_{j,t}|^p]$ (by convexity) for all $p \geq 2$. The bound (2.25) follows as in [Dedecker, 2001, proof of Corollary 3(b)], a method inspired by the proof of Theorem 6 in [Doukhan et al., 1984].

We first prove this corollary of Theorem 5. The constants are in no way optimal.

Corollary 6. *In the context of Theorem 5, for all $p \geq 2$,*

$$\mathbb{E}[|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]|^p] \leq \frac{3\pi}{2} \left[\left(\frac{15B \deg(\mathcal{G})}{2|\mathcal{A}|}\right)^p p^p + \left(\frac{32V \deg(\mathcal{G})}{|\mathcal{A}|}\right)^{p/2} p^{p/2} \right]. \quad (2.47)$$

Proof of Corollary 6. Fix arbitrarily $p \geq 2$. It is well known that

$$\mathbb{E}[U^p] = \int_0^\infty ps^{p-1}\mathbb{P}[U \geq s]ds$$

for any nonnegative random variable U . Let $r > 0$ be a constant that we will carefully choose later on. Set arbitrarily $s \geq 0$, define $m := s/r$, and introduce

$$\tilde{\mathcal{Z}}_m := |\mathcal{A}|^{-1} \sum_{\alpha \in \mathcal{A}} (\zeta_\alpha - \mathbb{E}[\zeta_\alpha]) \mathbf{1}\{|\zeta_\alpha - \mathbb{E}[\zeta_\alpha]| < m\}.$$

It holds that

$$\begin{aligned} \mathbb{P}(|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]| \geq s) &\leq \mathbb{P}[\mathcal{Z} - \mathbb{E}[\mathcal{Z}] \neq \tilde{\mathcal{Z}}_m] + \mathbb{P}[|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]| \geq s, \mathcal{Z} - \mathbb{E}[\mathcal{Z}] = \tilde{\mathcal{Z}}_m] \\ &\leq \mathbb{P}[r \max_{\alpha \in \mathcal{A}} |\zeta_\alpha - \mathbb{E}[\zeta_\alpha]| \geq s] + \mathbb{P}[|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]| \geq s, \mathcal{Z} - \mathbb{E}[\mathcal{Z}] = \tilde{\mathcal{Z}}_m] \\ &\leq \mathbb{P}[r \max_{\alpha \in \mathcal{A}} |\zeta_\alpha - \mathbb{E}[\zeta_\alpha]| \geq s] + \mathbb{P}[|\tilde{\mathcal{Z}}_m - \mathbb{E}[\tilde{\mathcal{Z}}_m]| \geq s - \mathbb{E}[\tilde{\mathcal{Z}}_m]] \end{aligned}$$

hence

$$\mathbb{E}[|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]|^p] \leq r^p \mathbb{E}[\max_{\alpha \in \mathcal{A}} |\zeta_\alpha - \mathbb{E}[\zeta_\alpha]|^p] + \int_0^\infty ps^{p-1} \mathbb{P}[|\tilde{\mathcal{Z}}_m - \mathbb{E}[\tilde{\mathcal{Z}}_m]| \geq s - \mathbb{E}[\tilde{\mathcal{Z}}_m]] ds. \quad (2.48)$$

We now note that

$$\begin{aligned} |\mathbb{E}[\tilde{\mathcal{Z}}_m]| &= |\mathbb{E}[\tilde{\mathcal{Z}}_m - (\mathcal{Z} - \mathbb{E}[\mathcal{Z}])]| = |\mathcal{A}|^{-1} \left| \mathbb{E} \left[\sum_{\alpha \in \mathcal{A}} (\zeta_\alpha - \mathbb{E}[\zeta_\alpha]) \mathbf{1}\{|\zeta_\alpha - \mathbb{E}[\zeta_\alpha]| \geq m\} \right] \right| \\ &\leq (m|\mathcal{A}|)^{-1} \sum_{\alpha \in \mathcal{A}} \text{Var}[\zeta_\alpha] = V/m. \end{aligned}$$

Therefore if $s \geq s_0 := \sqrt{2rV}$, then $s/2 \geq V/(s/r) = V/m$ hence $s - |\mathbb{E}[\tilde{\mathcal{Z}}_m]| \geq s/2$. In light of (2.46) and (2.48), the rightmost term in (2.48), say I_p , satisfies

$$\begin{aligned} I_p &\leq \int_0^{s_0} ps^{p-1} ds + \int_{s_0}^\infty ps^{p-1} \mathbb{P}[|\tilde{\mathcal{Z}}_m - \mathbb{E}[\tilde{\mathcal{Z}}_m]| \geq s/2] ds \\ &\leq s_0^p + 2 \int_{s_0}^\infty ps^{p-1} \exp\left(-\frac{|\mathcal{A}|\tilde{V}}{4m^2 \deg(\mathcal{G})} h\left(\frac{8ms/2}{5\tilde{V}}\right)\right) ds, \end{aligned} \quad (2.49)$$

where

$$\begin{aligned} \tilde{V} &:= |\mathcal{A}|^{-1} \sum_{\alpha \in \mathcal{A}} \text{Var}[(\zeta_\alpha - \mathbb{E}[\zeta_\alpha]) \mathbf{1}\{|\zeta_\alpha - \mathbb{E}[\zeta_\alpha]| \leq m\}] \\ &\leq |\mathcal{A}|^{-1} \sum_{\alpha \in \mathcal{A}} \mathbb{E}[(\zeta_\alpha - \mathbb{E}[\zeta_\alpha])^2 \mathbf{1}\{|\zeta_\alpha - \mathbb{E}[\zeta_\alpha]| \leq m\}] \\ &\leq |\mathcal{A}|^{-1} \sum_{\alpha \in \mathcal{A}} \mathbb{E}[(\zeta_\alpha - \mathbb{E}[\zeta_\alpha])^2] = V. \end{aligned}$$

Because $h(u) \geq \frac{u}{2} \log(1+u)$ for all $u \geq 0$, (2.49) yields

$$\begin{aligned} I_p &\leq s_0^p + 2 \int_{s_0}^{\infty} p s^{p-1} \exp\left(-\frac{|\mathcal{A}|s}{10m \deg(\mathcal{G})} \log\left(1 + \frac{4ms}{5\tilde{V}}\right)\right) ds \\ &= s_0^p + 2 \int_{s_0}^{\infty} p s^{p-1} \exp\left(-\frac{|\mathcal{A}|r}{10 \deg(\mathcal{G})} \log\left(1 + \frac{4s^2}{5r\tilde{V}}\right)\right) ds. \end{aligned} \quad (2.50)$$

If $u := s/(5r\tilde{V}/4)^{1/2}$, then $s^{p-1} \leq (5r\tilde{V}/4)^{(p-1)/2} (1+u^2)^{(p-1)/2}$. A change of variable and the bound $\tilde{V} \leq V$ thus imply that the rightmost term in (2.50) is smaller than

$$2p \left(\frac{5rV}{4}\right)^{p/2} \int_0^{\infty} (1+u^2)^{(p-1)/2-r|\mathcal{A}|/(10 \deg(\mathcal{G}))} du.$$

We now choose $r := 5(p+1) \deg(\mathcal{G})/|\mathcal{A}|$ to guarantee the convergence of the above integral, to $\pi/2$, and conclude that (2.48) and (2.50) imply

$$\begin{aligned} \mathbb{E}[|\mathcal{Z} - \mathbb{E}[\mathcal{Z}]|^p] &\leq r^p \mathbb{E}[\max_{\alpha \in \mathcal{A}} |\zeta_{\alpha} - \mathbb{E}[\zeta_{\alpha}]|^p] + \pi(rV)^{p/2} (2^{p/2} + p(5/4)^{p/2}) \\ &\leq (rB)^p + \pi(p+1)(2rV)^{p/2}. \end{aligned} \quad (2.51)$$

Finally, since $(p+1)/p \leq 3/2$ and $p^{2/p} \leq e^{2/e} \approx 2.61$, we can simplify (2.51) to (2.47), thus completing the proof of Corollary 6. \square

Fix arbitrarily $j \in \llbracket J \rrbracket$, $t \geq 1$, $V > 0$, and $p \geq 2$. To save space introduce, for each $\tau \in \llbracket t \rrbracket$,

$$\mathcal{Z}_{j,\tau} := \Delta^{\circ} \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_{\tau}, \bar{Z}_{\tau}) - \mathbb{E} \left[\Delta^{\circ} \bar{\ell}(\theta_{j,\tau-1})(\bar{O}_{\tau}, \bar{Z}_{\tau}) \middle| \bar{Z}_{\tau}, F_{\tau-1} \right].$$

In view of **A1**, **A2**, **A3** and **A5**, Corollary 6 applies and guarantees that almost surely, for all $\tau \in \llbracket t \rrbracket$,

$$\begin{aligned} \mathbb{E} \left[|\mathcal{Z}_{j,\tau}|^p \middle| \bar{Z}_{\tau}, F_{\tau-1} \right] \mathbf{1}\{\text{var}_{j,\tau} \leq V\} &\leq \frac{3\pi}{2} \left[\left(\frac{15b_2 \deg(\mathcal{G})}{2|\mathcal{A}|} \right)^p p^p \right. \\ &\quad \left. + \left(\frac{32V \deg(\mathcal{G})}{|\mathcal{A}|} \right)^{p/2} p^{p/2} \right] \mathbf{1}\{\text{var}_{j,\tau} \leq V\} \\ &\leq \frac{3\pi}{2} \left[\left(\frac{15b_2 \deg(\mathcal{G})}{2|\mathcal{A}|} \right)^p p^p + \left(\frac{32V \deg(\mathcal{G})}{|\mathcal{A}|} \right)^{p/2} p^{p/2} \right]. \end{aligned} \quad (2.52)$$

It is now easy to show that $\widetilde{\text{var}}_{t,j} \leq v_2$ almost surely (see (2.23) for the definition of v_2). By **A5**, it holds that $\text{var}_{j,\tau} \leq v_1$ almost surely for each $\tau \in \llbracket t \rrbracket$, hence

$$\widetilde{\text{var}}_{j,t} = \frac{1}{t} \sum_{\tau=1}^t \mathbb{E} \left[(\mathcal{Z}_{j,\tau})^2 \middle| \bar{Z}_{\tau}, F_{\tau-1} \right] = \frac{1}{t} \sum_{\tau=1}^t \mathbb{E} \left[(\mathcal{Z}_{j,\tau})^2 \middle| \bar{Z}_{\tau}, F_{\tau-1} \right] \mathbf{1}\{\text{var}_{j,\tau} \leq v_1\} \leq v_2$$

because of (2.52) with $p = 2$.

We now turn to the proof of (2.25). In view of (2.52), by convexity of $u \mapsto |u|^p$, it holds that

$$\begin{aligned} \mathbb{E} \left[\left| \widehat{H}_{j,t} - \widetilde{H}_{j,t} \right|^p \mathbf{1}\{\widetilde{\mathcal{F}}_V\} \right] &\leq \frac{1}{t} \sum_{\tau=1}^t \mathbb{E} \left[|\mathcal{Z}_{j,\tau}|^p \mathbf{1}\{\widetilde{\mathcal{F}}_V\} \right] \leq \frac{1}{t} \sum_{\tau=1}^t \mathbb{E} \left[|\mathcal{Z}_{j,\tau}|^p \mathbf{1}\{\text{var}_{j,\tau} \leq V\} \right] \\ &= \frac{1}{t} \sum_{\tau=1}^t \mathbb{E} \left[\mathbb{E} \left[|\mathcal{Z}_{j,\tau}|^p \mid \bar{Z}_\tau, F_{\tau-1} \right] \mathbf{1}\{\text{var}_{j,\tau} \leq V\} \right] \\ &\leq \frac{3\pi}{2} \left[\left(\frac{15b_2 \deg(\mathcal{G})}{2|\mathcal{A}|} \right)^p p^p + \left(\frac{32V \deg(\mathcal{G})}{|\mathcal{A}|} \right)^{p/2} p^{p/2} \right]. \end{aligned} \quad (2.53)$$

Therefore Markov's inequality implies that, for all $x > 0$,

$$\begin{aligned} \mathbb{P} \left[\left| \widehat{H}_{j,t} - \widetilde{H}_{j,t} \right| \geq x, \widetilde{\mathcal{F}}_V \right] &\leq \mathbb{E} \left[x^{-p} \left| \widehat{H}_{j,t} - \widetilde{H}_{j,t} \right|^p \mathbf{1}\{\widetilde{\mathcal{F}}_V\} \right] \\ &\leq \frac{3\pi}{2} \left(\frac{15b_2 \deg(\mathcal{G}) p/2 + \sqrt{32|\mathcal{A}|V \deg(\mathcal{G}) p}}{x|\mathcal{A}|} \right)^p. \end{aligned} \quad (2.54)$$

By the technical Lemma 7, there exists $p_x > 0$ such that

$$\begin{aligned} x|\mathcal{A}| &= 15eb_2 \deg(\mathcal{G}) p_x/2 + \sqrt{32e^2 |\mathcal{A}| V \deg(\mathcal{G}) p_x}, \quad \text{and} \\ p_x \geq q_x &:= (x|\mathcal{A}|)^2 \left(32e^2 |\mathcal{A}| V \deg(\mathcal{G}) + 15eb_2 \deg(\mathcal{G}) x|\mathcal{A}| \right)^{-1} \\ &= x^2 |\mathcal{A}| \left(32e^2 V \deg(\mathcal{G}) + 15eb_2 \deg(\mathcal{G}) x \right)^{-1}. \end{aligned}$$

If $q_x \geq 2$, then p_x is a valid choice for p in (2.54). This choice yields the inequality

$$\mathbb{P} \left[\left| \widehat{H}_{j,t} - \widetilde{H}_{j,t} \right| \geq x, \widetilde{\mathcal{F}}_V \right] \leq \frac{3\pi}{2} \exp(-p_x) \leq \frac{3\pi}{2} \exp(-q_x) \leq \exp(2 - q_x).$$

Otherwise, $\mathbb{P} \left[\left| \widehat{H}_{j,t} - \widetilde{H}_{j,t} \right| \geq x, \widetilde{\mathcal{F}}_V \right] \leq \exp(2 - q_x)$ holds trivially. This completes the proof of Theorem 3. \square

Lemma 7. *For any $a, b, c > 0$, there exists $p > 0$ such that $c = b\sqrt{p} + ap$. Moreover, $c^2 \leq (b^2 + 2ac)p$.*

Proof of Lemma 7. The quadratic equation $c = bX + aX^2$ has a positive solution, so there does exist $p > 0$ such that $c = b\sqrt{p} + ap$. Moreover, $c^2/p = b^2 + 2ab\sqrt{p} + a^2p$ on the one hand and $2ac = 2ab\sqrt{p} + 2a^2p \geq 2ab\sqrt{p} + a^2p$ on the other hand, implying that $c^2/p \leq b^2 + 2ac$. This completes the proof. \square

2.8 Appendix: a classical strong convexity argument

Suppose that Θ is convex, and that the loss function $\ell : \Theta \rightarrow \mathbb{R}^{\mathcal{O}}$ is a_1 -Lipschitz,

$$|\ell(\theta_1) - \ell(\theta_2)| \leq a_1 |\theta_1 - \theta_2| \quad (2.55)$$

and a_2 -strongly convex: for all $s \in [0, 1]$ and $\theta_1, \theta_2 \in \Theta$,

$$\ell(s\theta_1 + (1-s)\theta_2) - \frac{a_2}{2}(s\theta_1 + (1-s)\theta_2)^2 \leq s \left(\ell(\theta_1) - \left(\frac{a_2}{2}\theta_1\right)^2 \right) + (1-s) \left(\ell(\theta_2) - \left(\frac{a_2}{2}\theta_2\right)^2 \right)$$

(both inequalities above are understood pointwise). Then the modulus of continuity of ℓ is lower-bounded by $\rho \mapsto \frac{a_2}{8}\rho^2$ in the sense that, for all $\theta_1, \theta_2 \in \Theta$,

$$\frac{1}{2}(\ell(\theta_1) + \ell(\theta_2)) - \ell\left(\frac{1}{2}(\theta_1 + \theta_2)\right) \geq \frac{a_2}{8}(\theta_1 - \theta_2)^2 \quad (2.56)$$

(pointwise). Let P be a law on \mathcal{O} such that $P\ell(\theta)$ is well defined for all $\theta \in \Theta$, where we note $Pf := \int f dP$. Choose $\theta^\circ \in \Theta$ such that $P\ell(\theta^\circ) \leq P\ell(\theta)$ for all $\theta \in \Theta$. Then, for all $\theta \in \Theta$,

$$\begin{aligned} \frac{1}{2}P(\ell(\theta) + \ell(\theta^\circ)) &\geq P\ell\left(\frac{1}{2}(\theta + \theta^\circ)\right) + \frac{a_2}{8}P(\theta - \theta^\circ)^2 \\ &\geq P\ell(\theta^\circ) + \frac{a_2}{8}P(\theta - \theta^\circ)^2 \\ &\geq P\ell(\theta^\circ) + \frac{a_2}{8a_1^2}P(\ell(\theta) - \ell(\theta^\circ))^2, \end{aligned}$$

where the first inequality follows from (2.56), the second holds by convexity of Θ and choice of θ° , and the third one follows from (2.55). Therefore,

$$P(\ell(\theta^\circ) - \ell(\theta))^2 \leq \frac{4a_1^2}{a_2}P(\ell(\theta) - \ell(\theta^\circ)),$$

which concludes the argument.

Chapitre 3

L'anticipation des communes demanderessees de la reconnaissance de l'état de catastrophe naturelle par Super Learning

Ce chapitre est consacré à l'estimation des demandes de reconnaissance de l'état de catastrophe naturelle. Pour ce faire, une instance spécifique de l'algorithme présenté et étudié dans le chapitre précédent a été déployée. Comme dans le chapitre précédent, nous modélisons les données comme une série temporelle pour laquelle à chaque pas de temps $t \in \mathbb{N}^*$, l'observation O_t est elle-même composée d'une collection d'observations spatialement dépendantes $O_{\alpha,t}$, où α appartient à l'ensemble des communes de France métropolitaine \mathcal{A} . Chaque observation $O_{\alpha,t}$ se décompose en un couple $(\zeta_{\alpha,t}, X_{\alpha,t})$, où $\zeta_{\alpha,t}$ vaut 1 si la commune α a formulé une demande de reconnaissance l'année t et 0 sinon, et $X_{\alpha,t}$ est un ensemble de covariables constituant une description de cette même commune ainsi que de l'événement sécheresse ayant eu lieu au cours de l'année t . Notre objectif est d'estimer l'espérance conditionnelle θ^* de $\zeta_{\alpha,t}$ sachant $X_{\alpha,t}$. Une première section présentera le contexte et les enjeux de l'estimation des probabilités de demande de reconnaissance. Par la suite, les données mises à profit seront présentées, puis une instance de l'OSASSL sera déployée pour réaliser les prédictions. Dans une dernière section, une approche alternative de l'estimation des probabilités de demande de reconnaissance sera proposée dans le but d'améliorer les performances obtenues. Dans ce contexte le transfer learning sera appliqué à l'OSASSL.

Nous présentons dans ce chapitre des travaux postérieurs à ceux exposés dans les Chapitres 2 et 4. Cette entorse à la chronologie est justifiée par le souhait de fluidifier la narration.

Ces travaux n'ont pas encore fait l'objet d'une prépublication.

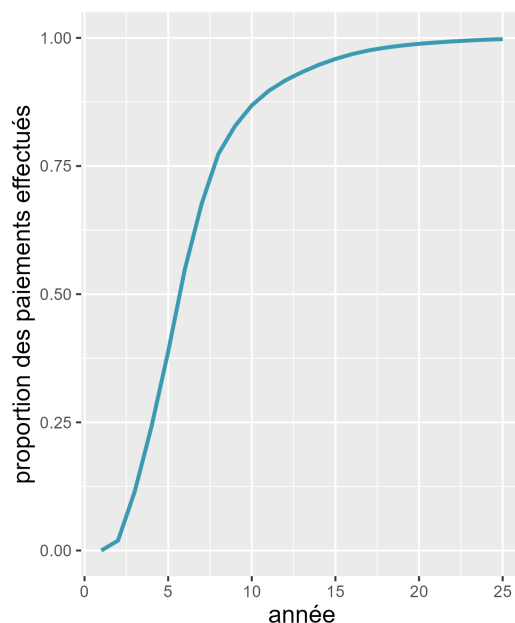


Figure 3.1: Rythme de paiement des sinistres liés à la sécheresse RGA effectués par CCR. Sept ans et demi après la survenance d'un événement sécheresse, 75% de sinistres ont été payés en moyenne.

3.1 Éléments sur la cinétique du péril sécheresse et estimation des demandes de reconnaissance

3.1.1 La sécheresse : un péril à déroulement long

La sécheresse géotechnique est provoquée par le mouvement des argiles présents dans le sol et provoque des dommages importants sur le bâti, principalement sur les maisons dont les fondations sont moins profondes que celles d'un immeuble ou d'un bâtiment industriel. Ce péril est notamment caractérisé par le cadencement particulièrement long des différentes étapes d'un sinistre. En effet, plusieurs mois peuvent être nécessaires à la constatation de fissures liées au phénomène. Par ailleurs, les assurés ne sont indemnisés qu'en cas de reconnaissance de l'état de catastrophe naturelle pour la commune où se situe le bien impacté, or la commission interministérielle traite les dossiers de demande de reconnaissance au second trimestre de l'année suivant la survenance de l'événement. À titre d'exemple, les premières demandes de reconnaissance relatives à l'événement de sécheresse de 2021 ont été traitées lors de la commission interministérielle du 14 juin 2022. Le cas échéant, il est également nécessaire qu'un expert d'assurance s'assure que le sinistre est bien lié à l'épisode de sécheresse pour lequel la commune est reconnue en état de catastrophe naturelle. Sans cela, le dossier serait classé sans-suite, et ne donnerait lieu à aucune indemnisation (tant de l'assuré, en provenance de l'assureur, que de l'assureur,

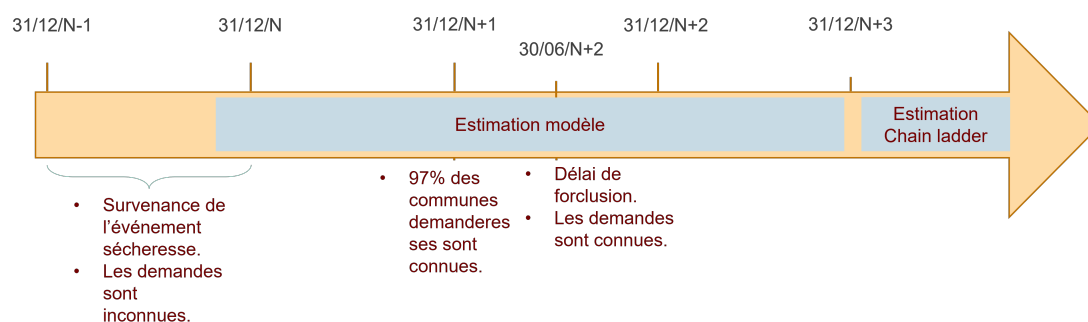


Figure 3.2: Chronologie des méthodes de prédiction déployées par CCR et évolution des communes demanderesses connues.

en provenance de CCR). Il peut s'avérer nécessaire d'attendre la consolidation du sinistre ou de réaliser une étude de sols pour compléter l'analyse de l'expert. Enfin, une fissure causée par la sécheresse RGA peut se rouvrir par la suite et certains sinistres font l'objet de recours.

En définitive, la cinétique du développement des dommages provoqués par le phénomène RGA, la nécessité d'attendre la décision de la commission interministérielle, les tensions sur la demande envers les experts et les entreprises d'études de sols expliquent que la sécheresse est qualifiée de péril à déroulement long. D'après la Figure 3.1, la sécheresse ne donne lieu à aucun paiement par CCR la première année. Pour un événement sécheresse donné, seulement 2% des paiements ont lieu la deuxième année. Finalement, plus de 20 années sont nécessaires en moyenne pour que l'ensemble des paiements liés à la survenance d'un événement sécheresse soient effectués.

3.1.2 Chronologie des méthodes d'estimation du coût d'un événement sécheresse

Des méthodes actuarielles existent pour l'estimation des dommages. À titre d'exemple, CCR a recours à la méthode du Chain ladder [Denuit and Charpentier, 2004] pour l'estimation des dommages liés aux anciens événements sécheresse. Cette méthode est basée sur l'extrapolation de la sinistralité survenue renseignée par les cédantes dans leur comptabilité qu'elles communiquent à CCR. Pour pouvoir déployer ce calcul, l'information comptable doit donc être disponible.

Dans la pratique, du fait du déroulement long du péril sécheresse, l'information comptable n'est disponible que tardivement et nécessite du temps avant de constituer une information robuste pouvant être extrapolée. Aussi, la méthode du Chain ladder n'est utilisée pour l'estimation des dommages qu'à partir de la 4^e année après la survenance d'un événement sécheresse. Pour cette raison, CCR a recours à des approches reposant sur des données extra-comptables telles que les modèles catastrophe pour réaliser les estimations des événements sécheresse liés aux 4 dernières années. Les modèles développés par CCR permettent donc de disposer d'une estimation des enjeux financiers liés à la

Date de valeur	Événement sécheresse	Contribution
31/12/2021	2018	40%
	2019	24%
	2020	33%
	2021	4%
31/12/2022	2019	19%
	2020	24%
	2021	2%
	2022	54%

Table 3.1: Contribution des événements sécheresse aux enjeux financiers générés par les 4 derniers événements.

survenance d'un événement sécheresse ayant eu lieu au cours des dernières années, ce avant que l'information comptable ne soit disponible ou stabilisée pour permettre le déploiement de méthodes actuarielles de type Chain ladder. Comme mentionné en Section 1.1.4, l'estimation des demandes de reconnaissance de l'état de catastrophe naturelle au titre de la sécheresse constitue une étape pour l'estimation du coût de chaque commune à l'aide du modèle catastrophe. Si les communes disposent de 18 mois après la survenance d'un événement sécheresse pour formuler une telle demande, on constate sur l'historique qu'au bout d'un an, 97% des communes demanderesses ont déjà réalisé cette démarche. La Figure 3.2 représente simultanément la connaissance progressive des communes demanderesses au cours du temps, ainsi que l'évolution des méthodes déployées par CCR pour l'estimation des dommages liés à un événement sécheresse. Bien que les estimations soient régulièrement mises à jour, les échéances représentées sur cette figure correspondent à la clôture des comptes de CCR. Lors de cet exercice, il est nécessaire de calculer le montant de sinistralité lié aux événements sécheresse afin de constituer des provisions, c'est-à-dire mettre de côté une partie de la prime collectée auprès des cédantes afin d'être en mesure de payer les sinistres lorsque ceux-ci seront présentés à CCR.

3.1.3 Les enjeux de l'anticipation des communes demanderesses dans le cadre du provisionnement

Si les différentes utilisations des modèles catastrophe ont été présentées en Section 1.1.3.2, les calculs réalisés dans le cadre de la clôture (correspondant au provisionnement) sont d'une importance majeure pour CCR.

La Table 3.1 présente les contributions respectives des événements aux enjeux financiers générés par les 4 derniers événements sécheresse au 31 décembre 2021 et au 31 décembre 2022. L'événement sécheresse de l'exercice courant est finalement le seul pour lequel l'estimation des demandes de reconnaissance est réellement nécessaire, puisqu'au bout d'un an 97% des demandes sont effectuées. Les clôtures des comptes au 31 décembre 2021 et au 31 décembre 2022 représentent des situations bien différentes. Dans le premier cas, l'événement lié à l'année en cours (2021) n'a pas un impact financier significatif par

rapport aux 3 autres événements (respectivement 4% et 96%). Ainsi, la précision des estimations des demandes de reconnaissance n'apparaît pas ici comme critique. À l'inverse, la précision de l'estimation des dommages pour les événements sécheresse de 2018 à 2020 pour lesquels les demandes sont connues est un enjeu important. Dans le second cas, l'événement sécheresse lié à l'année en cours (2022) a un impact financier majeur (54%). Ici, la précision de l'estimation des demandes de reconnaissance est capitale, mais la précision de l'estimation des dommages pour l'événement sécheresse de 2022 l'est tout autant du fait des enjeux financiers importants qu'il représente. Par ailleurs, l'estimation des dommages liés aux événements sécheresse des années 2019 à 2021, pour lesquels les demandes sont connues, reste un enjeu non négligeable (46% des dommages sur les 4 dernières années).

En définitive, dans le cadre de la clôture des comptes, l'estimation des demandes impacte uniquement l'estimation de l'événement lié à l'année en cours lorsque l'estimation des dommages s'avère elle indispensable pour l'ensemble des estimations. Pourtant, l'estimation des demandes de reconnaissance demeure un enjeu fort pour CCR dans la mesure où (i) la sinistralité liée à la sécheresse de l'exercice courant peut représenter des montants importants (par exemple en 2022) et (ii) notamment du fait de la dimension spatiale du phénomène étudié, les estimations des dommages présentent une sensibilité importante à l'estimation des demandes (nombre de demandes et localisation).

3.2 Les données

Le jeu de données exploité pour l'estimation des probabilités de demande de reconnaissance présente de nombreuses similitudes avec le jeu de données utilisé dans le chapitre précédent pour l'estimation des dommages liés à un événement sécheresse. Il existe néanmoins plusieurs différences entre ces deux jeux de données. En effet, de nouvelles observations et de nouvelles covariables ont été ajoutées, tandis que certaines covariables ont été supprimées.

La variation du nombre d'observations disponibles : Le nombre d'observations exploitées pour l'estimation des probabilités de demande de reconnaissance n'est pas le même que pour l'estimation des dommages pour les trois raisons suivantes :

L'ajout d'événements de sécheresse : Dans le chapitre précédent, les années postérieures à 2017 n'étaient pas considérées pour l'estimation des dommages car le montant de ces événements n'est pas encore stabilisé. En revanche, du fait du délai de forclusion imposé aux communes pour le dépôt de leur dossier de demande de reconnaissance, les demandes relatives aux événements sécheresse sont bien connues jusqu'à l'événement 2021. Le jeu de données a donc été enrichi des observations correspondant aux événements sécheresse des années 2018 à 2021.

La description de l'ensemble des communes de France métropolitaine : Dans ce nouveau jeu de données, l'ensemble des communes de France métropolitaine sont décrites pour chaque année. Dans le chapitre précédent, seules

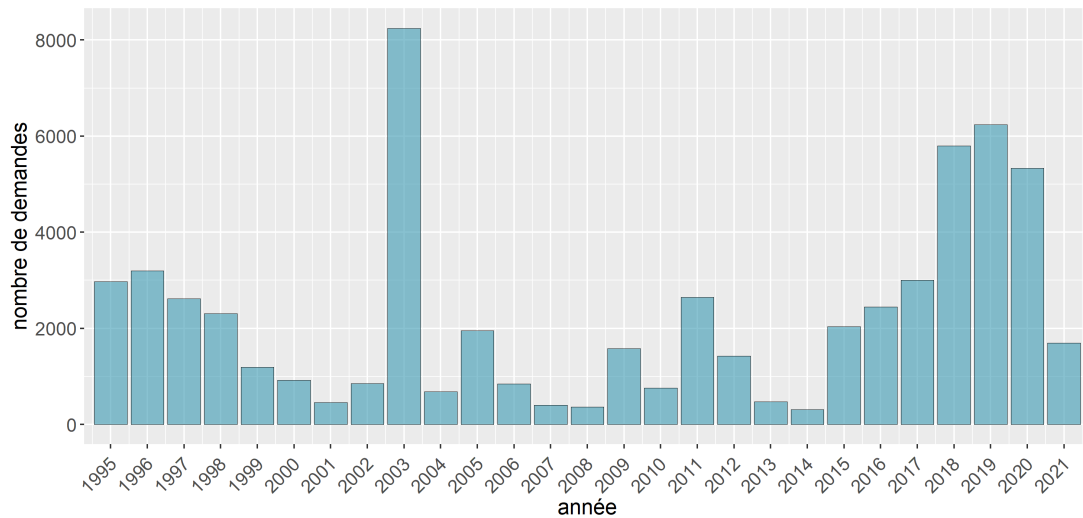


Figure 3.3: Nombre de demandes de reconnaissance de l'état de catastrophe naturelle au titre de la sécheresse RGA par année.

les communes ayant fait l'objet d'une reconnaissance de l'état de catastrophe naturelle étaient considérées.

La Figure 3.3 constitue une première restitution issue de l'exploitation du jeu de données constitué pour l'estimation des demandes de reconnaissance. On remarque que le nombre de demandes de reconnaissance annuel est très variable, allant de 313 en 2014 à 8 246 en 2003. Cette variabilité s'explique par la survenance d'événements sécheresse plus ou moins sévères selon les années. Les dernières années sont marquées par un nombre important de demandes, notamment en 2018, 2019 et 2020. Enfin, le jeu de données fait apparaître un taux de demandes (nombre de demandes moyen rapporté au nombre de communes) de 6,2%. Ainsi, l'estimation des probabilités de demande de reconnaissance correspond à un problème de classification dans le cas où la variable réponse présente un déséquilibre de classes.

La maîtrise de la volumétrie : Puisque l'ensemble des communes de France métropolitaine sont à présent décrites pour chaque année dans le jeu de données, la volumétrie est ici bien plus importante. Sur la période 1995-2021, le jeu de données comprend 978 620 observations. Afin de réduire la volumétrie, et donc les temps de calcul, nous utiliserons les données relatives aux années 2010-2021 dans le cadre de cette étude. Sur cette période, le nombre d'observations est de 430 129.

L'ajout de nouvelles covariables : Une première covariable a été ajoutée, indiquant si la commune est éligible ou non aux critères de reconnaissance de l'état de catastrophe naturelle. Logiquement, une covariable binaire a également été ajoutée afin d'identifier les communes demanderesse ou non.

Pour chaque commune et chaque année, le nombre de demandes et de reconnaissances de l'état de catastrophe naturelle dont a bénéficié la commune depuis 1990 et sur les 5 années précédentes permet de rendre compte de la dynamique temporelle des demandes. Ces covariables permettent également de représenter les habitudes des communes concernant la formulation des demandes de reconnaissance, ainsi que la connaissance ou non du régime d'indemnisation des catastrophes naturelles. Quatre autres covariables expriment ces mêmes quantités en taux, en les rapportant au nombre d'événements sécheresse observés.

Par la suite, 3 covariables renseignent le nombre de demandes de reconnaissance formulées par une commune mais rejetées, la dernière année (aucun ou un rejet), au cours des deux dernières années (0, 1 ou 2 rejets) et au cours des 5 dernières années (entre 0 et 5 rejets). Ces covariables liées à l'historique des demandes sont importantes notamment du fait de la dimension comportementale des demandes. En effet, certaines communes formulent une demande chaque année, et nous établissons à l'aide de notre jeu de données qu'une commune ayant fait l'objet d'un rejet d'une demande de reconnaissance de l'état de catastrophe naturelle l'année précédente a 7,8 fois plus de chances de reformuler une demande de reconnaissance l'année en cours.

La suppression de certaines covariables : Les ensembles de covariables suivants ont été supprimés du jeu de données :

- les variables relatives au coût, s'agissant ici d'estimer les demandes de reconnaissance ;
- les variables liées à la végétation, dans le but de maîtriser la volumétrie de notre jeu de données. Par ailleurs, ces variables apparaissaient secondaires pour l'estimation des dommages (voir Section 2.4.4) ;
- les variables liées à l'historique du SWI, également dans le but de maîtriser la volumétrie de notre jeu de données. Dans ce jeu de données, la dynamique temporelle est encodée dans les variables liées à l'historique des demandes, des reconnaissances et des rejets des demandes de reconnaissance. Les 36 valeurs de SWI de l'année courante ont également été écartées. Les variables correspondant à des résumés de celles-ci, telles que la valeur minimale du SWI, ont été conservées ;
- les variables correspondant aux 3 fonctions de répartition et aux moyennes pondérées par les valeurs assurées décrites en Section 2.2.3.3.

Le jeu de données ainsi obtenu est composé de 430 129 observations et 59 variables.

3.3 L'OSASSL pour l'estimation des demandes de reconnaissance

Une première approche pour l'estimation des demandes de reconnaissance consiste naturellement à transposer les travaux réalisés pour l'estimation des dommages à ce nouveau

problème. Plus précisément, il s’agit d’estimer les communes demanderesses de l’état de catastrophe naturelle à l’aide de l’algorithme présenté et étudié dans le chapitre précédent. Pour ce faire nous construisons une collection de J meta-algorithmes $\hat{\theta}_1, \dots, \hat{\theta}_J$, où chacun d’entre eux est un One Step Ahead Super Learner (OSASSL) s’appuyant sur une collection de K algorithmes fondamentaux $\hat{\mathcal{L}}_1, \dots, \hat{\mathcal{L}}_K$. Enfin, un Super Learner, lui-même un OSASSL, réalise la tâche de combiner les prédictions des meta-algorithmes. Nommé overarching Super Learner, ce dernier sélectionne le meilleur meta-algorithme à l’issue d’une validation croisée séquentielle (on parle alors d’overarching Super Learner discret) ou bien réalise une combinaison linéaire convexe de leurs prédictions (on parle alors d’overarching Super Learner continu). Contrairement à la prédiction des dommages qui correspondait à une tâche de régression, l’OSASSL est ici mis à contribution dans le cadre d’une tâche de classification.

3.3.1 L’architecture retenue et l’implémentation

3.3.1.1 Les algorithmes fondamentaux et les meta-algorithmes

A nouveau, l’implémentation de l’OSASSL a été réalisée à l’aide du langage de programmation R [R Core Team, 2022]. Une collection de $K = 5$ algorithmes fondamentaux a été implémentée. Cette collection comprend :

- une régression logistique (`stats::glm`, [R Core Team, 2022]),
- une forêt aléatoire (`ranger::ranger`, [Wright and Ziegler, 2017]),
- un gradient boosting avec booster de type arbre de classification (`xgboost::xgb.train`, [Chen et al., 2021]),
- un gradient boosting avec booster de type régression logistique (`xgboost::xgb.train`, [Chen et al., 2021]),
- un réseau de neurones profond (`keras::fit`, [Allaire and Chollet, 2021]).

Une seconde collection de $J = 7$ meta-algorithmes a été implémentée. Cette collection comprend :

- une régression logistique (`stats::glm`, [R Core Team, 2022]),
- une forêt aléatoire (`ranger::ranger`, [Wright and Ziegler, 2017]),
- un gradient boosting avec booster de type arbre de classification (`xgboost::xgb.train`, [Chen et al., 2021]),
- un arbre de classification (`rpart::rpart`, [Therneau and Atkinson, 2019]),
- un réseau de neurones profond (`keras::fit`, [Allaire and Chollet, 2021]),
- un Super Learner sélectionnant le meilleur algorithme fondamental (`SuperLearner::SuperLearner`, [Polley et al., 2021]),
- un second Super Learner réalisant une combinaison linéaire convexe des prédictions des algorithmes fondamentaux (`SuperLearner::SuperLearner`, [Polley et al., 2021]).

La Figure 3.11 en annexe représente l’architecture de cet OSASSL.

3.3.1.2 Les métriques

Le suivi de plusieurs métriques permet d'apprécier les performances obtenues et de comparer différents modèles entre eux. Dans notre étude, nous choisissons comme fonction de perte l'erreur quadratique. Dans le cadre de la validation croisée séquentielle, l'erreur quadratique moyenne se calcule de la façon suivante, par exemple pour l'algorithme indexé par j et au pas de temps $t \in \mathbb{N}^*$:

$$\frac{1}{t|\mathcal{A}|} \sum_{\tau=1}^t \sum_{\alpha \in \mathcal{A}} (\zeta_{\alpha,\tau} - \theta_{j,\tau-1}(X_{\alpha,\tau}))^2.$$

L'écart quadratique moyen est à la fois la quantité que cherche à minimiser l'overarching Super Learner, et à la fois une métrique que nous retenons pour le suivi des performances du modèle. En plus de l'écart quadratique moyen, nous retenons la logloss moyenne comme seconde métrique, que nous noterons Logloss dans la suite de ce document. Dans le cadre de la validation croisée séquentielle, la Logloss se calcule de la façon suivante, par exemple pour l'algorithme indexé par j et au pas de temps $t \in \mathbb{N}^*$:

$$-\frac{1}{t|\mathcal{A}|} \sum_{\tau=1}^t \sum_{\alpha \in \mathcal{A}} (\zeta_{\alpha,\tau} \ln(\theta_{j,\tau-1}(X_{\alpha,\tau})) + (1 - \zeta_{\alpha,\tau}) \ln(1 - \theta_{j,\tau-1}(X_{\alpha,\tau}))).$$

Le choix de l'erreur quadratique moyenne est motivé par l'emploi fréquent de cette métrique, sous l'appellation Brier score, dans la littérature traitant des prédictions atmosphériques et météorologiques pour l'évaluation de prédictions de probabilités [voir à titre d'exemple Jolliffe and Stephenson, 2012, page 157]. La Logloss est une métrique plus commune dans le cas d'une classification, notamment en machine learning [voir Zhang and Yang, 2004, Bennett, 2003, Zadrozny and Elkan, 2001].

3.3.1.3 L'apprentissage séquentiel

Les trois couches de l'algorithme sont entraînées séquentiellement. Les 5 algorithmes $\widehat{\mathcal{L}}_1, \dots, \widehat{\mathcal{L}}_5$ constituant la première couche sont entraînés séquentiellement à partir des données $\bar{O}_{2010}, \dots, \bar{O}_{2020}$. Ainsi pour chaque $t \in \llbracket 2011, 2021 \rrbracket$ et chaque $k \in \llbracket 1, 5 \rrbracket$, l'algorithme $\widehat{\mathcal{L}}_k$ produit un estimateur $\ell_{k,t-1}$ entraîné sur les données $\bar{O}_{2010}, \dots, \bar{O}_{t-1}$. Cet estimateur est utilisé pour réaliser des prédictions $\ell_{k,t-1}(X_{\alpha,t})$ pour les données \bar{O}_t . Les prédictions obtenues alimentent les 7 meta-algorithmes $\widehat{\theta}_1, \dots, \widehat{\theta}_7$ constituant la seconde couche. Pour chaque $t \in \llbracket 2012, 2021 \rrbracket$ et chaque $j \in \llbracket 1, 7 \rrbracket$, l'algorithme $\widehat{\theta}_j$ produit un estimateur $\theta_{j,t-1}$ entraîné sur les données $(\bar{O}_{2011}, \{\ell_{k,2010}(X_{\alpha,2011}) : k \in \llbracket 1, 5 \rrbracket, \alpha \in \mathcal{A}\}), \dots, (\bar{O}_{t-1}, \{\ell_{k,t-2}(X_{\alpha,t-1}) : k \in \llbracket 1, 5 \rrbracket, \alpha \in \mathcal{A}\})$. Cet estimateur est utilisé pour réaliser des prédictions $\theta_{j,t-1}(X_{\alpha,t})$ pour les données \bar{O}_t . Pour finir, ces prédictions alimentent l'overarching Super Learner discret ou continu. Pour chaque $t \in \llbracket 2013, 2021 \rrbracket$, l'overarching Super Learner est entraîné sur les données $\{(\theta_{j,2011}(X_{\alpha,2012}), \zeta_{\alpha,2012}) : j \in \llbracket 1, 7 \rrbracket, \alpha \in \mathcal{A}\}, \dots, \{(\theta_{j,t-2}(X_{\alpha,t-1}), \zeta_{\alpha,t-1}) : j \in \llbracket 1, 7 \rrbracket, \alpha \in \mathcal{A}\}$ et réalise les prédictions pour les données $\{\theta_{j,t-1}(X_{\alpha,t}) : j \in \llbracket 1, 7 \rrbracket, \alpha \in \mathcal{A}\}$.

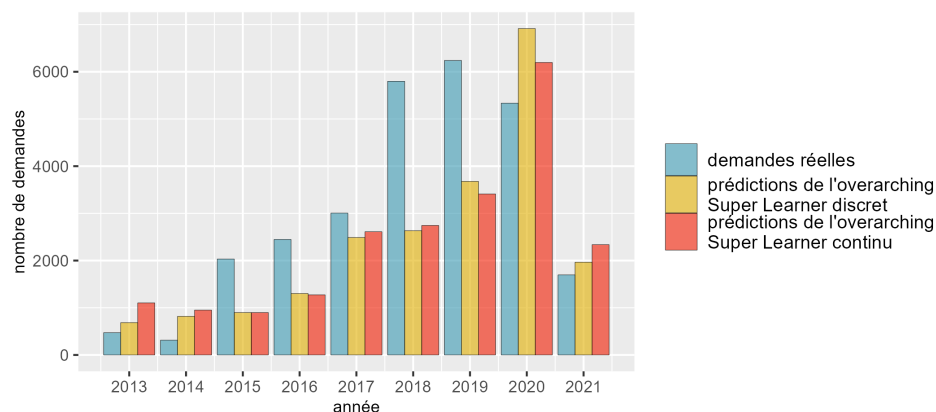


Figure 3.4: Somme des probabilités de demande prédites annuellement par l'overarching discret et continu, comparaison aux demandes réelles.

Algorithme	MSE	Logloss
Overarching discret	0,0563	0,1930
Overarching continu	0,0564	0,1932

Table 3.2: Synthèse des métriques calculées pour les prédictions du Super Learner overarching discret et continu.

Le temps de calcul correspondant à la mise en œuvre de l'ensemble de la procédure d'apprentissage de l'algorithme et à la réalisation des prédictions est d'une heure.

3.3.2 Les résultats

Les prédictions de l'overarching Super Learner ne sont disponibles qu'à partir de l'année 2013, dans la mesure où les prédictions des meta-algorithmes nécessaires à son apprentissage ne sont disponibles qu'à partir de l'année 2012. La Figure 3.4 présente une première restitution des résultats obtenus. En présentant la somme des probabilités de demande prédites annuellement par l'overarching Super Learner et la somme des demandes réelles, elle permet notamment de rendre compte de la proximité des prédictions réalisées par les deux versions de l'overarching. Malheureusement, l'écart entre la somme des prévisions de l'overarching Super Learner et la somme des demandes réelles est dans l'ensemble important, notamment pour les événements sécheresse des années 2018 et 2019. La Table 3.2 présente les métriques choisies calculées pour l'ensemble des prédictions et pour les deux versions de l'overarching Super Learner. Elle confirme la proximité entre ces deux versions de l'algorithme, et indique un léger avantage pour la version discrète de l'overarching, pour lequel ressort à la fois une meilleure MSE (0,0563 contre 0,0564) et un meilleur Logloss (0,1930 contre 0,1932). Enfin, la Figure 3.5 représente une cartographie des prédictions des probabilités de demande de reconnaissance obtenues à l'aide de

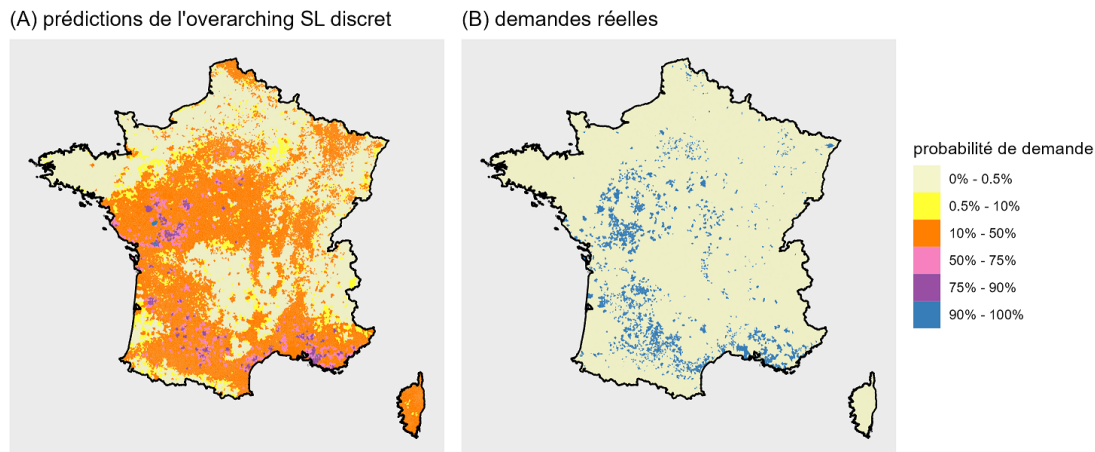


Figure 3.5: Cartographie des prédictions des probabilités de demande pour l'événement sécheresse de 2021 réalisées par le Super Learner overarching discret.

l'overarching Super Learner discret pour l'événement sécheresse de l'année 2021. À la vue de cette représentation il apparaît que les communes pour lesquelles l'overarching discret estime une forte probabilité de demande sont bien situées dans des zones présentant une forte concentration de demandes. Cependant, pour de nombreuses communes non demanderesses, le modèle estime tout de même une faible probabilité de demande. À l'exception de la Corse, l'empreinte spatiale des demandes est correctement anticipée, mais le ciblage est très approximatif. La suite de cette étude vise à améliorer ces résultats.

3.4 Un nouvel angle de modélisation pour l'anticipation des demandes de reconnaissance

Dans cette section, nous proposons une nouvelle modélisation du problème étudié. Visant à améliorer les résultats obtenus précédemment, cette modélisation innovante est rendue possible par l'exploitation d'une nouvelle source de données en provenance de la commission interministérielle.

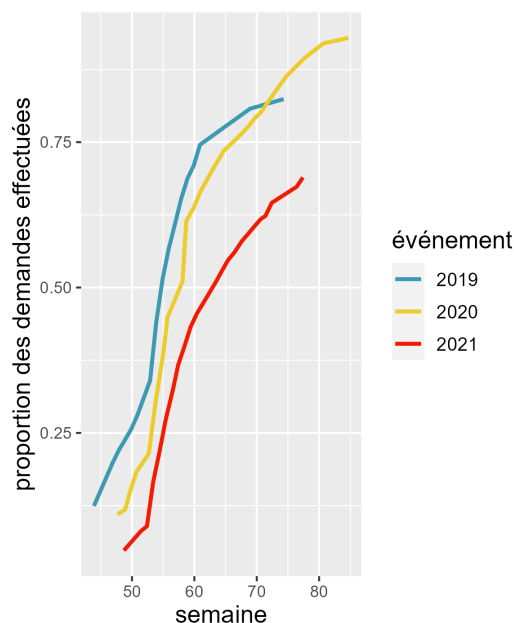


Figure 3.6: Rythme de dépôt des demandes de reconnaissance de l'état de catastrophe naturelle au titre de la sécheresse RGA observé sur les événements 2019 à 2021.

3.4.1 L'estimation dynamique des demandes de reconnaissance

3.4.1.1 Une nouvelle source de données

Dans le cadre du régime d'indemnisation des catastrophes naturelles, les communes impactées par un événement naturel sont amenées à formuler une demande de reconnaissance de l'état de catastrophe naturelle. Ces demandes sont ensuite traitées par une commission interministérielle qui prononce un avis favorable ou non sur la base de critères préétablis. Du fait de la cinétique du péril sécheresse, les demandes relatives à ces événements sont traitées au deuxième trimestre de l'année suivant leur survenance. Cependant, les communes peuvent formuler leur demande bien avant le début du traitement des dossiers par la commission, notamment au cours de l'année de survenance de l'événement sécheresse.

En tant que secrétaire de la commission interministérielle, CCR reçoit des fichiers constituant un état des lieux de la réception des dossiers de demande de reconnaissance au titre de la sécheresse. Ces fichiers, reçus à partir du mois d'octobre ou novembre de l'année de l'événement, contiennent notamment les informations suivantes :

- la date de valeur de l'état des lieux, c'est-à-dire la date à laquelle la commission interministérielle a réalisé le décompte des communes demanderesse ;
- la liste des communes ayant formulé une demande de reconnaissance, à cette date de valeur ou bien précédemment ;

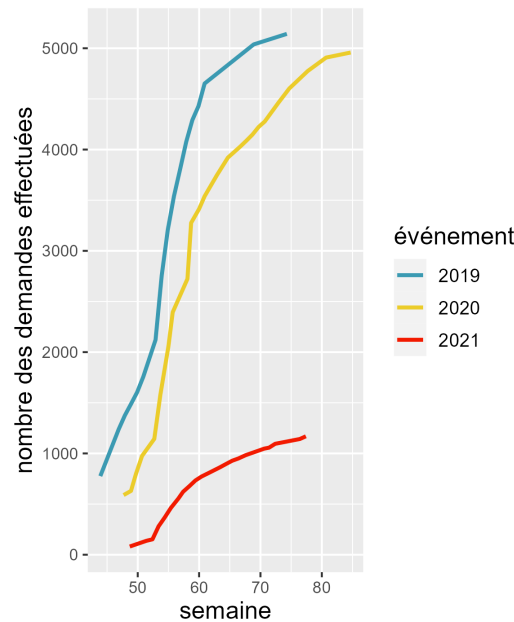


Figure 3.7: *Évolution du nombre de demandes de reconnaissance de l'état de catastrophe naturelle au titre de la sécheresse RGA observé sur les événements 2019 à 2021.*

- pour chacune d'elles, le code INSEE et le nombre de bâtiments touchés. Cette dernière information est renseignée de manière déclarative par les maires des communes.

Ces fichiers permettent d'étudier la dynamique des dépôts des demandes de reconnaissance. Dans la mesure où le processus permettant la communication de ces fichiers n'a été stabilisé qu'à partir de l'année 2019, et dans la mesure où le délai de forclusion n'est pas atteint pour l'événement sécheresse de 2022 au moment de l'écriture de ce manuscrit, seuls les événements sécheresse 2019 à 2021 peuvent être étudiés sous cet angle.

La Figure 3.6 présente l'évolution des demandes de reconnaissance pour les événements sécheresse de 2019, 2020 et 2021 en fonction de la date de valeur. On observe notamment que le rythme de dépôt des dossiers diffère selon les événements. À titre d'exemple, les dossiers liés à la sécheresse 2021 ont été déposés plus tardivement, probablement en lien avec la faible intensité de l'événement. À l'inverse, les demandes de reconnaissance liées aux événements 2019 et 2020, beaucoup plus marqués en termes de dommages, ont été déposées beaucoup plus rapidement. Ces trajectoires ont en commun une forte accélération des demandes entre les semaines 55 et 65.

La Figure 3.7 représente l'évolution hebdomadaire du stock de demandes de reconnaissance pour les événements sécheresse 2019 à 2021. Précédemment dans la Section 3.3.2, la somme des probabilités de demande prédites pour l'événement sécheresse de 2019 était de l'ordre de 3 500, contre un nombre de demandes réelles supérieur à 6 000. Pourtant,

la Figure 3.7 indique que le stock de demandes liées à cet événement a dépassé 3500 dès la semaine 57. La prise en compte du nombre de demandes en stock apparaît alors être une information pouvant améliorer les prédictions. La connaissance de la localisation des communes constituant ce stock présente également un intérêt. Pour une date de valeur donnée, une commune n'ayant pas formulé de demande de reconnaissance sera plus susceptible de le faire si elle est entourée de plusieurs communes ayant déjà formulé une telle demande. En effet, en plus du fait que des communes puissent s'influencer entre elles, cette proximité avec des communes demanderesse indique une localisation dans une zone impactée par l'événement.

En définitive, les fichiers en provenance de la commission interministérielle renseignent CCR sur le contexte dans lequel se trouvent certaines communes, et plus globalement sur la dynamique temporelle et spatiale des demandes de reconnaissance liées à l'événement pour lequel il s'agit d'estimer les probabilités de demande de reconnaissance in fine des communes. L'approche déployée dans la section précédente ne permet pas de prendre en compte ces éléments, pourtant de nature à améliorer les performances des prédictions. Aussi, l'exploitation de ces informations nécessite de s'appuyer sur une modélisation alternative.

3.4.1.2 Une modélisation originale

Les données en provenance de la commission interministérielle décrites précédemment permettent de modéliser différemment le problème que nous souhaitons résoudre. Nous introduisons les notations qui seront utilisées par la suite. Pour chaque année $t \in \{2019, 2020, 2021\}$, \mathcal{U}_t correspond à l'ensemble des numéros de semaine pour lesquelles nous bénéficions d'un fichier de données en provenance de la commission interministérielle relatif à l'événement sécheresse de l'année t . Ces ensembles sont les suivants :

$$\mathcal{U}_{2019} = \{44, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 58, 59, 60, 61, 69, 75\},$$

$$\mathcal{U}_{2020} = \{48, 49, 50, 51, 53, 54, 55, 56, 58, 59, 60, 61, 63, 65, 67, 68, 69, \\ 70, 71, 73, 75, 78, 81, 85\},$$

$$\mathcal{U}_{2021} = \{49, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 64, 65, 66, 67, 71, 72, 73, 77, 78\}.$$

Les semaines indexées au-delà de 52 correspondent à des observations de l'événement sécheresse l'année suivant sa survenance. On remarque que les premières informations sont communiquées à CCR entre fin octobre (semaine 44 de 2019) et début décembre (semaine 49 de 2021). Par ailleurs, ces données ne couvrent pas l'ensemble de la période précédant le délai de forclusion, correspondant environ à la semaine 130.

Également, pour chaque $t \in \{2019, 2020, 2021\}$, $\alpha \in \mathcal{A}$ et $u \in \mathcal{U}_t$:

- $\zeta_{\alpha,t}$ vaut 1 si la commune α a formulé une demande de reconnaissance au titre de l'événement t , 0 sinon. La quantité $\sum_{\alpha \in \mathcal{A}} \zeta_{\alpha,t}$ correspond au nombre de demandes de reconnaissance formulées au titre de l'événement t . Il s'agit de la variable réponse.
- $\zeta_{\alpha,t,u}$ vaut 1 si lors de la semaine u la commune α a déjà formulé une demande de reconnaissance, cette demande pouvant avoir été formulée au cours de la semaine

- u ou bien précédemment. En d'autres termes, cette variable aléatoire vaut 1 si la commune α appartient au stock de demandes de la semaine u pour l'événement t .
- $\mathcal{A}_{t,u}^- \subset \mathcal{A}$ est l'ensemble des communes $\alpha \in \mathcal{A}$ n'ayant pas formulé de demande de reconnaissance la semaine u ou précédemment, au titre de l'événement t : $\mathcal{A}_{t,u}^- = \{\alpha \in \mathcal{A} : \zeta_{\alpha,t,u} = 0\}$. Ces communes seront désignées par la suite par le terme "communes négatives", dans le sens où elles n'ont pas formulé de demande de reconnaissance à ce stade. Parmi elles, certaines formuleront par la suite une demande de reconnaissance alors que d'autres n'entreprendront pas cette démarche.
 - $\mathcal{A}_{t,u}^+ = \overline{\mathcal{A}_{t,u}^-} \subset \mathcal{A}$ est l'ensemble des communes $\alpha \in \mathcal{A}$ ayant formulé une demande de reconnaissance la semaine u ou précédemment au titre de l'événement t : $\mathcal{A}_{t,u}^+ = \{\alpha \in \mathcal{A} : \zeta_{\alpha,t,u} = 1\}$. Ces communes correspondent au stock de demandes et seront désignées par la suite par le terme "communes positives", dans le sens où elles ont déjà formulé une demande de reconnaissance.
 - $X_{\alpha,t,u} \in \mathcal{X}$ correspond au vecteur de covariables décrivant la commune α l'année t en semaine u .

Dans cette nouvelle approche, l'objectif est de prédire pour chaque semaine de l'événement 2021 la probabilité de demande de reconnaissance des communes négatives. Ces prédictions seront réalisées à l'aide des données relatives aux exercices précédents. Les données d'apprentissage sont les suivantes :

$$\{(X_{\alpha,t,u}, \zeta_{\alpha,t}) : t \in \{2019, 2020\}, u \in \mathcal{U}_t, \alpha \in \mathcal{A}_{t,u}^-\},$$

Les prédictions sont réalisées pour les données suivantes :

$$\{X_{\alpha,2021,u} : u \in \mathcal{U}_{2021}, \alpha \in \mathcal{A}_{2021,u}^-\}.$$

Cette nouvelle modélisation est appelée par la suite l'approche dynamique, dans le sens où les estimations des probabilités de demande sont réalisées de façon dynamique à la réception de chaque fichier de données en provenance de la commission interministérielle. Par opposition, l'approche annuelle mise en œuvre en Section 3.3 sera appelée par la suite l'approche statique.

Via l'introduction de nouvelles covariables, l'approche dynamique présentera l'avantage de mettre à profit les informations disponibles dans les états communiqués à CCR par la commission interministérielle.

3.4.1.3 Le nouveau jeu de données

La modélisation présentée précédemment nécessite la création d'un nouveau jeu de données. Celui-ci comprendra notamment des covariables résumant les informations transmises par la commission interministérielle.

Pour chaque événement sécheresse $t \in \{2019, 2020, 2021\}$ le processus suivant a été appliqué itérativement à chaque fichier en provenance de la commission interministérielle de date de valeur $u \in \mathcal{U}_t$:

- récupérer le jeu de données relatif à l'événement t issu de l'approche statique ;
- ajouter les covariables suivantes représentant un résumé du contexte de la commune ainsi que de la dynamique temporelle et spatiale de l'événement à différentes échelles :
 - **à l'échelle locale** : le nombre de bâtiments touchés par l'événement sécheresse recensés dans les communes voisines, la moyenne du rapport entre ce nombre de bâtiments touchés et le nombre de maisons de chaque commune voisine, le nombre de communes positives parmi les communes voisines, le nombre de communes voisines positives pour la première fois, le nombre de communes voisines positives pour la première fois sur les 5 dernières années, la moyenne du taux de demandes des communes positives voisines correspondant au rapport entre le nombre de demandes historiques effectuées depuis 1990 et le nombre d'événements sécheresse, la moyenne du taux de demandes des communes positives voisines calculé sur les 5 dernières années, le rapport entre le taux de demandes de la commune rapporté à la date de valeur ;
 - **à l'échelle départementale** : le nombre de bâtiments touchés par l'événement dans le département, la proportion moyenne de bâtiments touchés dans les communes du département, le nombre de communes positives pour la première fois, le nombre de communes positives pour la première fois sur les 5 dernières années, le taux de demandes moyen, le taux de demandes moyen calculé sur les 5 dernières années ;
 - **à l'échelle nationale** : la date de valeur u , le nombre de communes positives, le rapport entre le logarithme du nombre de communes positives et la date de valeur ;
- supprimer l'ensemble des communes positives.

Ainsi pour chaque fichier en provenance de la commission interministérielle, un jeu de données est créé comprenant l'ensemble des communes de France métropolitaine à l'exception de celles ayant formulé une demande de reconnaissance à la date de valeur du fichier. La concaténation de ces jeux de données intermédiaires permet d'obtenir un dernier jeu de données, composé d'un peu plus de 2 millions d'observations et de 76 variables. C'est ce jeu de données que nous exploiterons dans la suite de cette étude.

3.4.2 OSASSL et transfer learning⁵

3.4.2.1 Motivations

Dans cette section, nous avons proposé une nouvelle approche pour l'estimation des probabilités de demande de reconnaissance. Cette nouvelle approche présente notamment l'avantage de mettre à profit une nouvelle source de données riche et prometteuse. En contrepartie, cette approche amène à réduire considérablement la profondeur de l'historique que nous pouvons considérer. La Table 3.3 renseigne les quartiles et la moyenne de

5. Apprentissage par transfert en français. Nous conservons l'appellation anglaise dans ce manuscrit.

Événement sécheresse	min.	1 ^{er} qu.	median	moyenne	3 ^e qu.	max.
2019 à 2020	-0.39	0.18	0.20	0.20	0.24	0.88
2021	-0.22	0.21	0.26	0.29	0.35	0.95
1995 à 2021	-0.48	0.19	0.23	0.25	0.29	1.10

Table 3.3: *Résumé de la distribution du minimum du SWI des communes de France métropolitaine pour les événements sécheresse 2019 et 2020, pour l'événement sécheresse 2021 et pour les événements sécheresse de 1995 à 2021.*

la distribution du minimum du SWI des communes pour les événements sécheresse 2019 et 2020, qui sont les événements disponibles pour l'apprentissage, et pour l'événement sécheresse 2021, qui est l'événement réservé pour la validation. Enfin, la distribution du minimum du SWI des communes pour les événements sécheresses 1995 à 2021 est également résumée pour comparaison. Il apparaît que les événements sécheresse 2019 et 2020 ont été particulièrement marqués. En effet, les valeurs des 1^{er}, 2^e et 3^e quartiles de la distribution du minimum du SWI sont inférieures à ce qui a été observé sur la période 1995-2021. À l'inverse, l'événement sécheresse de 2021 a été particulièrement clément : les 1^{er}, 2^e et 3^e quartiles de la distribution du minimum du SWI des communes sont systématiquement supérieurs à ces mêmes statistiques observées sur la période 1995-2021. Ainsi, en réduisant significativement la profondeur de l'historique du jeu de données, l'approche dynamique de l'estimation des demandes de reconnaissance réduit également la diversité des événements sécheresse auxquels est exposé le modèle lors de l'apprentissage.

Dans ce contexte, le transfer learning, un procédé visant à améliorer les performances d'un algorithme en lui transférant des connaissances acquises dans un contexte connexe, sera mis à profit. L'objectif est de continuer à profiter des données antérieures à 2019 dans le cadre de l'approche dynamique de l'estimation des demandes de reconnaissance.

3.4.2.2 Transfer learning

Nous introduisons à présent les grands principes du transfer learning.

Définition 1 (Domaine). *Un domaine \mathcal{D} est un ensemble composé de deux éléments : le domaine de définition des covariables noté \mathcal{X} et la distribution des covariables notée \mathbb{P}_X . On note $\mathcal{D} = \{\mathcal{X}, \mathbb{P}_X\}$.*

Définition 2 (Tâche). *Pour un domaine $\mathcal{D} = \{\mathcal{X}, \mathbb{P}_X\}$ donné, une tâche \mathcal{T} correspond à un ensemble de deux éléments : l'espace des labels noté \mathcal{Y} et une fonction notée $f(\cdot)$. On note $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, où $f(\cdot)$ est une fonction que nous cherchons à apprendre pour la réalisation des prédictions à partir d'observations à valeurs dans \mathcal{X} . Dans le cas d'une classification, il peut par exemple s'agir de la fonction qui à toute loi pour (X, Y) associe la probabilité conditionnelle de $Y = 1$ sachant X .*

Une fois les notions de domaine et de tâche définies, nous pouvons définir formellement

le transfer learning. La définition suivante correspond au cas où il n'y a qu'un unique domaine source.

Définition 3 (Transfer learning). *Soient un domaine source \mathcal{D}_S , une tâche source \mathcal{T}_S , un domaine cible \mathcal{D}_T et une tâche cible \mathcal{T}_T . Le transfer learning correspond aux procédés permettant d'améliorer l'apprentissage de la fonction cible $f_T(\cdot)$ sur \mathcal{D}_T à l'aide des connaissances sur \mathcal{D}_S et \mathcal{T}_S , avec $\mathcal{D}_S \neq \mathcal{D}_T$ ou $\mathcal{T}_S \neq \mathcal{T}_T$.*

Il existe une certaine diversité parmi les situations pour lesquelles le transfer learning peut être mis à profit. Dans [Pan and Yang, 2010] et [Zhuang et al., 2021], une première catégorisation est proposée distinguant la disponibilité ou non des labels pour les domaines source ou cible. Si ces labels ne sont disponibles que pour le domaine source, on parle alors de transfer learning transductif. Les tâches \mathcal{T}_S et \mathcal{T}_T sont identiques et les domaines \mathcal{D}_S et \mathcal{D}_T différents. Si les données du domaine source et cible sont labellisées, on parle alors de transfert learning inductif. Les tâches \mathcal{T}_S et \mathcal{T}_T diffèrent et les domaines \mathcal{D}_S et \mathcal{D}_T sont identiques ou non. Enfin, on parle de transfer learning non supervisé si aucune donnée n'est labellisée. Les tâches \mathcal{T}_S et \mathcal{T}_T sont différentes et les domaines \mathcal{D}_S et \mathcal{D}_T sont identiques ou non.

Il est également possible de catégoriser les situations dans lesquelles le transfer learning peut être profitable en fonction du domaine de définition des covariables et de la variable réponse pour le domaine source et cible. Si les domaines de définition des covariables et de la variable réponse sont les mêmes entre le domaine source et le domaine cible, c'est-à-dire $\mathcal{X}_S = \mathcal{X}_T$ et $\mathcal{Y}_S = \mathcal{Y}_T$, on parle alors de transfer learning homogène. Dans le cas contraire, i.e. $\mathcal{X}_S \neq \mathcal{X}_T$ ou $\mathcal{Y}_S \neq \mathcal{Y}_T$, on parle alors de transfer learning hétérogène.

Plusieurs méthodes de transfer learning ont été développées afin de répondre aux différents contextes décrits précédemment. Dans [Pan and Yang, 2010] et [Zhuang et al., 2021], les méthodes de transfer learning sont classées en 4 grandes catégories : les méthodes fondées sur les instances, celles fondées sur la représentation des covariables, celles fondées sur le transfert de la connaissance de relations existantes dans le jeu de données et enfin celles fondées sur les paramètres des modèles.

Les méthodes fondées sur les instances. Ces méthodes de transfer learning reposent sur la pondération des observations, par exemple dans le but de tenir compte de dérives dans le temps au niveau d'une ou plusieurs covariables.

Les méthodes fondées sur la représentation des covariables. Ces méthodes de transfer learning reposent sur la transformation des covariables, par exemple dans le but de créer un nouvel espace dans lequel les covariables des domaines source et cible correspondraient.

Les méthodes fondées sur le transfert de la connaissance de relations. Ces méthodes reposent sur le réemploi dans le domaine cible des relations entre variables constatées dans le domaine source.

Les méthodes fondées sur les paramètres des modèles. Dans cette approche de transfer learning, il est supposé que les modèles dédiés aux tâches source et cible partagent certains paramètres. Le transfer learning basé sur les réseaux est une approche très populaire et adaptée aux réseaux de neurones [Tan et al., 2018, Zhuang et al., 2021]. Dans cette approche, un premier réseau de neurones est entraîné dans le domaine source. Ensuite, une partie de ce modèle est extraite et agrégée pour finalement correspondre aux premières couches d'un second réseau de neurones qui sera déployé dans le domaine cible. On parle alors d'extraction de covariables. Éventuellement, les poids des premières couches de ce deuxième réseau de neurones peuvent être légèrement ajustés (fine-tuning) pour mieux correspondre à la tâche cible. On parle alors d'initialisation de poids.

Dans le cadre de cette étude, la tâche source correspond à l'estimation annuelle des probabilités de demande de reconnaissance de l'état de catastrophe naturelle. La tâche cible correspond à l'estimation hebdomadaire de ces mêmes probabilités. Les domaines source et cible ne sont pas non plus identiques du fait de l'introduction de nouvelles covariables décrites en Section 3.4.1. Le contexte est donc celui du transfer learning inductif hétérogène. Par ailleurs, le modèle développé dans le cadre de cette étude présente des similitudes avec un réseau de neurones profond. À ce titre, nous proposons d'employer le transfer learning basé sur les réseaux tel que déployé dans [Pinto et al., 2022]. À travers l'extraction de covariables du domaine source, disposant d'une plus grande profondeur d'historique, ce procédé permettra l'enrichissement du jeu de données et à travers cela l'amélioration des performances de l'algorithme déployé pour la réalisation de la tâche cible.

3.4.3 Mise en œuvre

3.4.3.1 Les architectures retenues

L'approche dynamique de l'estimation des demandes de reconnaissance décrite précédemment repose sur une modélisation originale du problème permise par l'exploitation de nouvelles données en provenance de la commission interministérielle. Cependant, du fait de la jeunesse de ces données, cette approche permet seulement l'exploitation des données relatives aux événements de sécheresse 2019 à 2021. Cette restriction sur l'historique implique des contraintes sur le choix de l'architecture de l'OSASSL ayant abouties à la suppression de la couche des meta-algorithmes. En effet, chaque couche de l'OSASSL "coûte" une année du fait de la validation séquentielle. Dans le cas de l'OSASSL à deux couches proposé ici, les algorithmes fondamentaux débutent leur apprentissage sur les données de l'exercice sécheresse de 2019. Par la suite, l'overarching agrège les algorithmes fondamentaux sur la base des performances de ces derniers constatées sur les données de l'année 2020. Enfin, l'overarching est en mesure d'effectuer des prédictions pour l'événement sécheresse 2021, soit le dernier événement disponible dans cette approche. Il apparaît donc impossible de considérer une couche supplémentaire à ce stade.

Dans le cadre de l'approche dynamique de l'estimation des demandes de reconnaissance, deux instances de l'OSASSL ont été implémentées. La première reprend simplement

l'implémentation présentée en Section 3.3.1 pour l'estimation statique des demandes de reconnaissance, à laquelle la couche des meta-algorithmes a été supprimée. La seconde reprend cette même structure et met en œuvre le transfer learning. La couche des 5 algorithmes fondamentaux est alimentée par :

- La couche des meta-algorithmes de l'implémentation de l'OSASSL pour l'estimation statique des demandes de reconnaissance. Celle-ci est elle-même alimentée par une couche d'algorithmes fondamentaux, elle-même alimentée par la couche des données en entrée.
- Une seconde couche d'entrées comprenant les nouvelles covariables liées à l'approche dynamique décrites en Section 3.4.1.

La mise en œuvre du transfer learning basé sur les réseaux dans le cadre de l'OSASSL consiste à enrichir le jeu de données des prédictions de la couche des meta-algorithmes obtenues dans le cadre de l'approche statique de l'estimation des demandes de reconnaissance. La Figure 3.12 en annexe représente l'architecture de l'OSASSL bénéficiant du transfer learning.

3.4.3.2 L'apprentissage séquentiel

Dans le cadre de l'approche dynamique sans transfer learning, correspondant à la première instance de l'OSASSL décrite dans la section précédente, les 5 algorithmes fondamentaux notés $\widehat{\mathcal{L}}_1^D, \dots, \widehat{\mathcal{L}}_5^D$ sont entraînés séquentiellement à partir des données relatives aux événements sécheresse des années 2019 et 2020 : $\{(X_{\alpha,t,u}, \zeta_{\alpha,t}) : t \in \{2019, 2020\}, u \in \mathcal{U}_t, \alpha \in \mathcal{A}_{t,u}^-\}$. Ainsi pour chaque $t \in \{2020, 2021\}$ et chaque $k \in \llbracket 1, 5 \rrbracket$, l'algorithme $\widehat{\mathcal{L}}_k^D$ produit un estimateur $\ell_{k,t-1}^D$ entraîné sur les données $\{(X_{\alpha,\tau,u}, \zeta_{\alpha,\tau}) : \tau \in \llbracket 2019, t-1 \rrbracket, u \in \mathcal{U}_\tau, \alpha \in \mathcal{A}_{\tau,u}^-\}$. Cet estimateur est utilisé pour réaliser des prédictions $\ell_{k,t-1}^D(X_{\alpha,t,u})$ pour les données $\{(X_{\alpha,t,u}, \zeta_{\alpha,t}) : u \in \mathcal{U}_t, \alpha \in \mathcal{A}_{t,u}^-\}$. Les prédictions obtenues alimentent l'overarching Super Learner discret ou continu. L'overarching Super Learner est entraîné sur les données

$$\{(\ell_{k,2019}^D(X_{\alpha,2020,u}), \zeta_{\alpha,2020}) : u \in \mathcal{U}_{2020}, \alpha \in \mathcal{A}_{2020,u}^-, k \in \llbracket 1, 5 \rrbracket\}$$

et réalise les prédictions pour les données

$$\{(\ell_{k,2020}^D(X_{\alpha,2021,u}), \zeta_{\alpha,2021}) : u \in \mathcal{U}_{2021}, \alpha \in \mathcal{A}_{2021,u}^-, k \in \llbracket 1, 5 \rrbracket\}.$$

L'apprentissage de l'OSASSL dans le cadre de l'approche dynamique avec transfer learning, correspondant à la seconde instance de l'OSASSL décrite dans la section précédente, s'effectue de façon analogue. Les données mises à contribution par les algorithmes fondamentaux notés $\widehat{\mathcal{L}}_1^{TL}, \dots, \widehat{\mathcal{L}}_5^{TL}$ sont cependant enrichies des prédictions des 7 meta-algorithmes de l'OSASSL implémenté dans le cadre de l'approche statique (voir Section 3.3.1.3). Ainsi, pour chaque $t \in \{2020, 2021\}$ et chaque $k \in \llbracket 1, 5 \rrbracket$, l'algorithme

Approche	Algorithme	MSE	Logloss
Statique	Overarching SL discret	0,0353	0,1279
	Overarching SL continu	0,0358	0,1301
Dynamique	Overarching SL discret	0,0238	0,0945
	Overarching SL continu	0,0234	0,0919
Dynamique avec transfer learning	Overarching SL discret	0,0236	0,0938
	Overarching SL continu	0,0231	0,0907

Table 3.4: Synthèse des métriques MSE et Logloss calculées pour l'événement sécheresse 2021 dans le cadre de l'approche statique, dynamique et dynamique avec transfer learning

$\widehat{\mathcal{L}}_k^{TL}$ produit un estimateur $\ell_{k,t-1}^{TL}$ entraîné sur les données $\{(X_{\alpha,\tau,u}, \theta_{j,t-1}(X_{\alpha,\tau}), \zeta_{\alpha,\tau}) : \tau \in \llbracket 2019, t-1 \rrbracket, u \in \mathcal{U}_\tau, \alpha \in \mathcal{A}_{\tau,u}^-, j \in \llbracket 1, 7 \rrbracket\}$. Cet estimateur est utilisé pour réaliser des prédictions $\ell_{k,t-1}^{TL}(X_{\alpha,t,u})$ pour les données $\{(X_{\alpha,t,u}, \theta_{j,t-1}(X_{\alpha,t}), \zeta_{\alpha,t}) : u \in \mathcal{U}_t, \alpha \in \mathcal{A}_{t,u}^-, j \in \llbracket 1, 7 \rrbracket\}$. Les prédictions obtenues alimentent l'overarching Super Learner discret ou continu. L'overarching Super Learner est entraîné sur les données

$$\{(\ell_{k,2019}^{TL}(X_{\alpha,2020,u}), \zeta_{\alpha,2020}) : u \in \mathcal{U}_{2020}, \alpha \in \mathcal{A}_{2020,u}^-, k \in \llbracket 1, 5 \rrbracket\}$$

et réalise les prédictions pour les données

$$\{(\ell_{k,2020}^{TL}(X_{\alpha,2021,u}), \zeta_{\alpha,2021}) : u \in \mathcal{U}_{2021}, \alpha \in \mathcal{A}_{2021,u}^-, k \in \llbracket 1, 5 \rrbracket\}.$$

Dans le cas de l'approche dynamique, le temps de calcul correspondant à la mise en œuvre de l'ensemble de la procédure d'apprentissage de l'algorithme et à la réalisation des prédictions est de 50 minutes. Dans le cas de l'approche dynamique avec transfer learning, le temps de calcul est de 1 heure et 30 minutes. Si les prédictions des meta-algorithmes réalisées dans le cadre de l'approche statique ont été préalablement calculées, le temps de calcul est alors réduit à 30 minutes.

3.4.4 Les résultats

Au terme de l'apprentissage séquentiel, l'overarching Super Learner continu sans transfer learning pondère à 51% le meta-algorithme correspondant à un réseau de neurones profond et à 49% le meta-algorithme correspondant à un gradient boosting fondé sur des arbres. L'overarching Super Learner discret sans transfer learning a sélectionné le meta-algorithme fondé sur un réseau de neurones profond. L'overarching Super Learner continu avec transfer learning pondère lui à 49% le meta-algorithme correspondant à un réseau de neurones profond, à 45% le meta-algorithme correspondant à un gradient boosting fondé sur des arbres, et à 6% le meta-algorithme correspondant à une forêt aléatoire. L'overarching Super Learner discret avec transfer learning a sélectionné le meta-algorithme fondé sur un réseau de neurones profond.

Dans le but de comparer les résultats issus de l'approche dynamique à ceux de l'approche statique, la Table 3.4 représente les métriques considérées précédemment moyennées sur l'ensemble \mathcal{U}_{2021} des dates de valeur de l'événement sécheresse 2021. Ainsi, la MSE est calculée de la façon suivante pour les prédictions de l'événement 2021 :

$$\frac{1}{|\mathcal{U}_{2021}|} \sum_{u \in \mathcal{U}_{2021}} \frac{1}{|\mathcal{A}_{2021,u}^-|} \sum_{\alpha \in \mathcal{A}_{2021,u}^-} (\zeta_{\alpha,2021} - \widehat{oa}_{\alpha,2021,u})^2,$$

où $\widehat{oa}_{\alpha,2021,u}$ désigne les prédictions réalisées par l'overarching Super Learner pour la commune α , l'année 2021 et la semaine u . La Logloss est calculée de façon analogue :

$$\begin{aligned} \frac{1}{|\mathcal{U}_{2021}|} \sum_{u \in \mathcal{U}_{2021}} \frac{1}{|\mathcal{A}_{2021,u}^-|} \sum_{\alpha \in \mathcal{A}_{2021,u}^-} (\zeta_{\alpha,2021} \ln(\widehat{oa}_{\alpha,2021,u})) \\ + (1 - \zeta_{\alpha,2021}) \ln(1 - \widehat{oa}_{\alpha,2021,u}). \end{aligned}$$

La Table 3.4 recense les métriques obtenues pour l'approche statique, l'approche dynamique sans transfer learning ainsi que l'approche dynamique avec transfer learning. On observe que l'approche dynamique permet d'améliorer significativement les prédictions. La MSE de l'overarching Super Learner est de 0,0238 dans sa version discrète et de 0,0234 dans sa version continue, contre une MSE de 0,0353 pour l'overarching discret qui obtient les meilleures performances dans l'approche statique. L'analyse des Logloss calculées aboutie aux mêmes conclusions. Enfin, le transfer learning permet une légère amélioration des performances tant du point de vue de la MSE que de la Logloss. Ainsi, l'overarching Super Learner continu avec transfer learning permet d'obtenir les meilleurs résultats.

Pour des raisons de confidentialité, nous ne pouvons proposer dans ce manuscrit une comparaison détaillée des résultats obtenus dans cette étude et de ceux obtenus à l'aide du modèle actuellement utilisé par CCR. Néanmoins, nous rapportons une amélioration de la MSE supérieure à 20%. La Figure 3.8 représente une vue dynamique des performances de l'overarching Super Learner continu avec transfer learning. Pour chaque date de valeur, on observe une très forte proximité entre d'une part le stock de demandes augmenté de la somme des probabilités de demande estimées pour les communes négatives, et d'autre part le nombre de demandes réel correspondant au trait en pointillés. La courbe noire correspond à la MSE représentée sous la forme d'une fonction $u \mapsto \frac{1}{|\mathcal{A}_{2021,u}^-|} \sum_{\alpha \in \mathcal{A}_{2021,u}^-} (\zeta_{\alpha,2021} - \widehat{oa}_{\alpha,2021,u})^2$. La décroissance de cette courbe signifie qu'en même temps que les informations en provenance de la commission interministérielle deviennent de plus en plus riches au fur et à mesure des semaines, les performances des prédictions s'améliorent progressivement.

Les analyses des performances ne pouvant être réalisées que sur un seul événement, nous fournissons en Annexe 3.7.2 cette même figure produite pour l'événement sécheresse de 2020 sur la base des prédictions du meta-algorithme fondé sur un réseau de neurones profond, correspondant à l'algorithme sélectionné par l'overarching Super Learner discret avec transfer learning. Les résultats sont également encourageants.

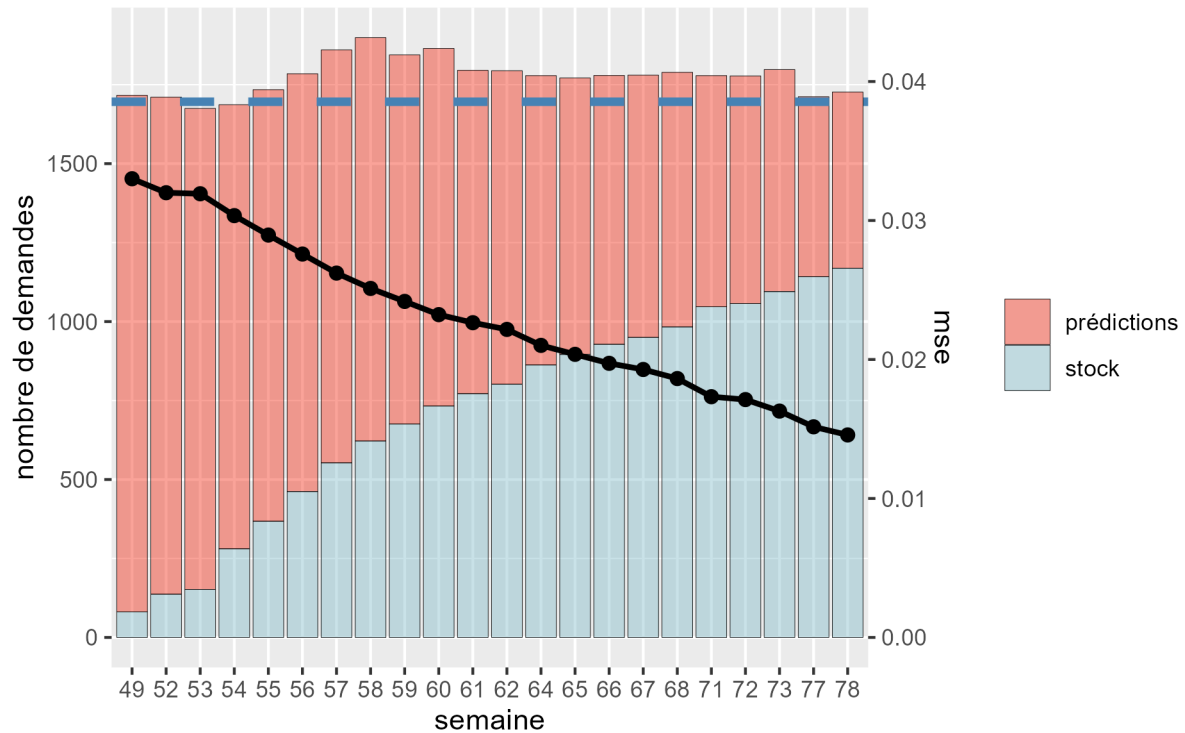


Figure 3.8: *MSE, stock de demandes et somme des prédictions des probabilités de demande de reconnaissance au titre de l'événement sécheresse 2021 pour chaque semaine étudiée. La MSE est représentée par la courbe noire correspondant à la fonction $u \mapsto \frac{1}{|\mathcal{A}_{2021,u}^-|} \sum_{\alpha \in \mathcal{A}_{2021,u}^-} (\zeta_{\alpha,2021} - \widehat{o\alpha}_{\alpha,2021,u})^2$. Le trait bleu en pointillés représente le nombre de demandes réelles.*

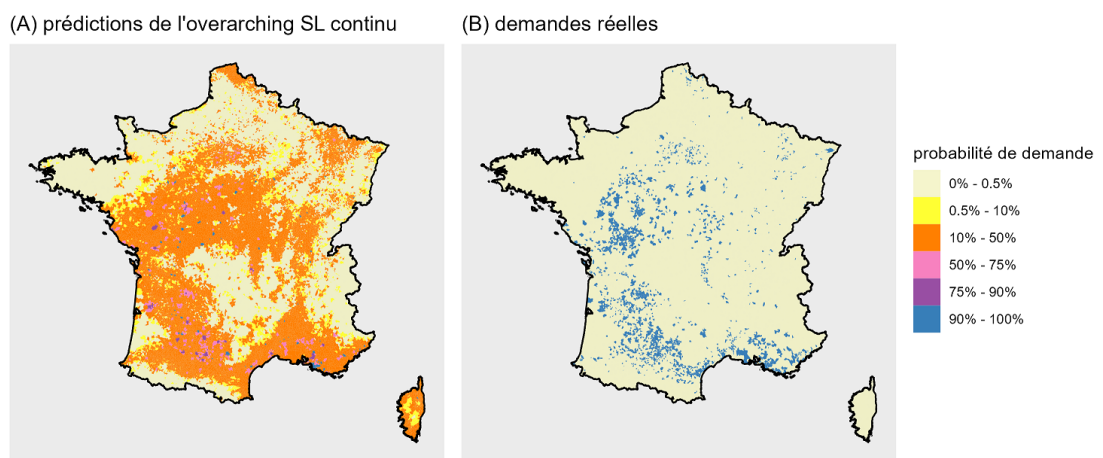


Figure 3.9: Cartographie des prédictions des probabilités de demande pour l'événement sécheresse de 2021 réalisées par l'overarching Super Learner continu avec transfer learning en semaine 49. Sur la carte de gauche, les communes appartenant au stock sont en bleu.

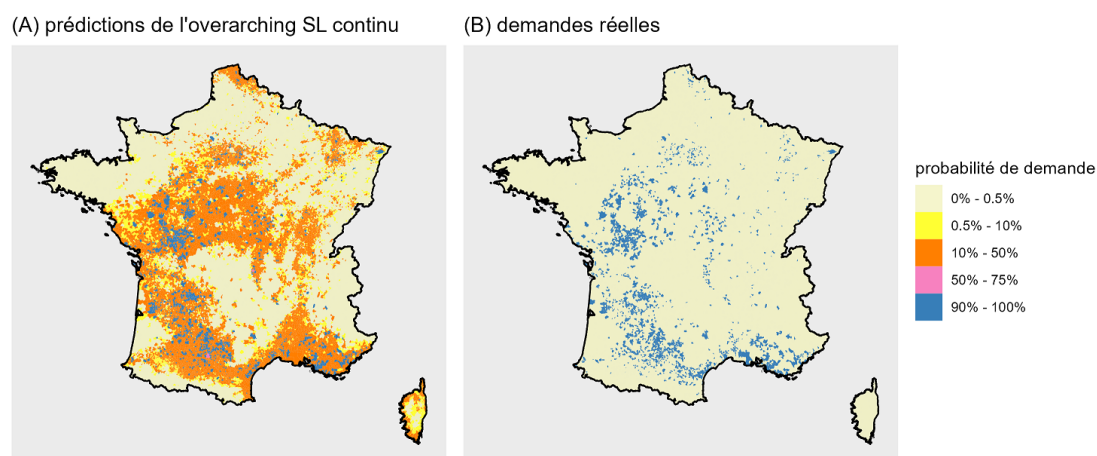


Figure 3.10: Cartographie des prédictions des probabilités de demande pour l'événement sécheresse de 2021 réalisées par l'overarching Super Learner continu avec transfer learning en semaine 78. Sur la carte de gauche, les communes appartenant au stock sont en bleu.

Les Figures 3.9 et 3.10 représentent une cartographie des prédictions des probabilités de demande de reconnaissance réalisées pour l'événement 2021 par l'overarching Super Learner continu dans le cadre de l'approche dynamique avec transfer learning, respectivement en semaine 49 et 78, comparées aux demandes réelles. La première cartographie illustre une première amélioration par rapport à l'approche statique puisqu'une partie moins importante du territoire se voit affecter une probabilité de demande de reconnaissance significative. La seconde cartographie fait apparaître une nette amélioration au fil des semaines, puisqu'en semaine 78 l'empreinte de l'événement est mieux cernée par le modèle. Ce constat est cohérent avec l'amélioration de la MSE au cours des semaines commentée précédemment. Le nord et l'est du territoire font néanmoins apparaître des concentrations de demandes de reconnaissance qui ne correspondent pas aux demandes réelles. Enfin, on note sur cette deuxième carte qu'aucune commune ne se voit affecter une probabilité comprise entre 75% et 90%.

3.5 Éléments conclusifs

Dans ce chapitre, une instance de l'OSASSL, l'algorithme développé et étudié dans le chapitre précédent dans le cadre de l'estimation des dommages, a été implémentée pour l'estimation des probabilités de demande de reconnaissance de l'état de catastrophe

naturelle au titre d'un événement de sécheresse RGA. Une approche identique a été entreprise : les prédictions ont été réalisées pour chaque événement de sécheresse et pour chaque commune. Le jeu de données a été adapté afin de correspondre à l'anticipation des probabilités de demande de reconnaissance. Du fait de la volumétrie, certaines variables ont été écartées et l'architecture de l'OSASSL a été simplifiée. Dans la mesure où les résultats obtenus n'ont pas apporté satisfaction, une approche alternative a été proposée. Cette nouvelle approche repose sur une source de données en provenance de la commission interministérielle, relativement récente et jusqu'à présent non-exploitée dans les modèles. Cette approche permet de réaliser des estimations quasi hebdomadaires et de plus en plus précises en tenant compte de l'augmentation de l'information au cours du temps. Cette approche dynamique réduit cependant la profondeur de l'historique considéré car fondé sur des données disponibles et fiables depuis 2019 seulement. Ainsi, le transfer learning basé sur les réseaux inspiré du deep learning et appliqué à l'OSASSL permet de capitaliser sur les sources de données anciennes et nouvelles à la fois. Nous avons montré que l'approche dynamique permet d'obtenir des prédictions significativement plus précises, le transfer learning permettant d'améliorer plus encore ces dernières. Enfin, les résultats obtenus constituent une amélioration par rapport au modèle actuellement déployé par CCR.

3.6 Perspectives de recherche

Bien que le transfer learning permette de bénéficier de l'information en provenance des anciennes et des nouvelles données, l'approche dynamique de l'estimation des probabilités de demande de reconnaissance conserve deux défauts liés au faible historique des données en provenance de la commission interministérielle :

- L'architecture de l'OSASSL demeure contrainte. Ainsi, la couche des meta-algorithmes a été supprimée ;
- La validation ne peut être effectuée que sur la base de l'exercice sécheresse de l'année 2021. Une validation reposant sur d'avantage d'événements gagnerait en robustesse.

La réalisation d'un bootstrap spatio-temporel permettrait le déploiement de l'approche dynamique sur l'ensemble de la période 1995-2021. Pour un événement t de la période 1995-2018 ($t \in \llbracket 1995, 2018 \rrbracket$), il s'agirait de générer des fichiers fictifs contenant les mêmes informations que ceux communiqués par la commission interministérielle. Le procédé pourrait être le suivant :

1. tirer aléatoirement une des 3 trajectoires de dépôt de dossiers de demande de reconnaissance $u \mapsto \frac{|\mathcal{A}_{\tau,u}^+|}{\sum_{\alpha \in \mathcal{A}} \zeta_{\alpha,\tau}}$ observées sur la période 2019-2021 ($\tau \in \llbracket 2019, 2021 \rrbracket$) et représentées en Figure 3.6 ;
2. pour chaque semaine $u \in \mathcal{U}_{\square}$, déduire le nombre de demandes $|\mathcal{A}_{t,u}^+|^*$ constituant le stock en appliquant le coefficient de la trajectoire retenue au nombre de demandes réelles de l'événement t : $|\mathcal{A}_{t,u}^+|^* = \sum_{\alpha \in \mathcal{A}} \zeta_{\alpha,t} \cdot \frac{|\mathcal{A}_{\tau,u}^+|}{\sum_{\alpha \in \mathcal{A}} \zeta_{\alpha,\tau}}$. Ensuite :

- (a) tirer aléatoirement parmi l'ensemble des communes demanderesse pour cet événement $\{\alpha \in \mathcal{A} : \zeta_{\alpha,t} = 1\}$ un nombre de communes égal au stock $|\mathcal{A}_{t,u}^+|^*$ calculé précédemment ;
- (b) renseigner le nombre de bâtiments touchés à l'aide des données de sinistres à l'adresse collectées par CCR en provenance de ses cédantes.

Cette approche par bootstrap, prometteuse, nécessiterait toutefois d'étudier les dynamiques spatiales et temporelles des dépôts des dossiers de demande de reconnaissance dans le but de générer des données fictives réalistes.

3.7 Annexe

3.7.1 Annexe A : représentations des architectures des OSASSL

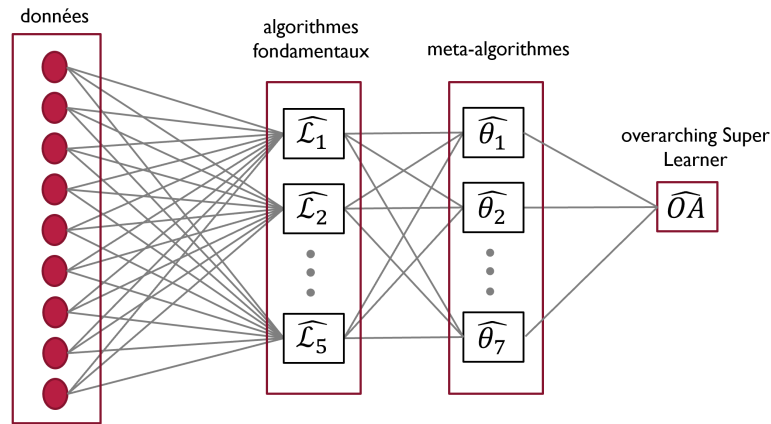


Figure 3.11: Représentation de l'architecture de l'OSASSL déployé pour l'estimation des demandes de reconnaissance.

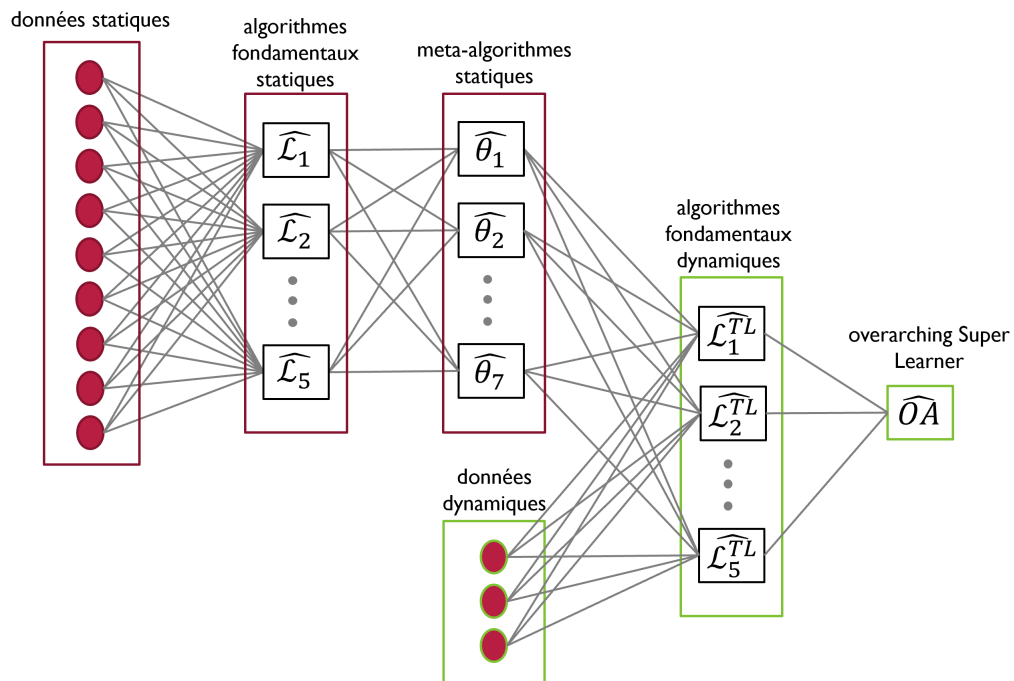


Figure 3.12: Représentation de l'architecture de l'OSASSL déployé pour l'estimation des demandes de reconnaissance dans le cadre de l'approche dynamique avec transfer learning.

3.7.2 Annexe B : prédictions des probabilités de demande de reconnaissance au titre de l'événement sécheresse de 2020 réalisées par le meta-algorithme fondé sur un réseau de neurones profond

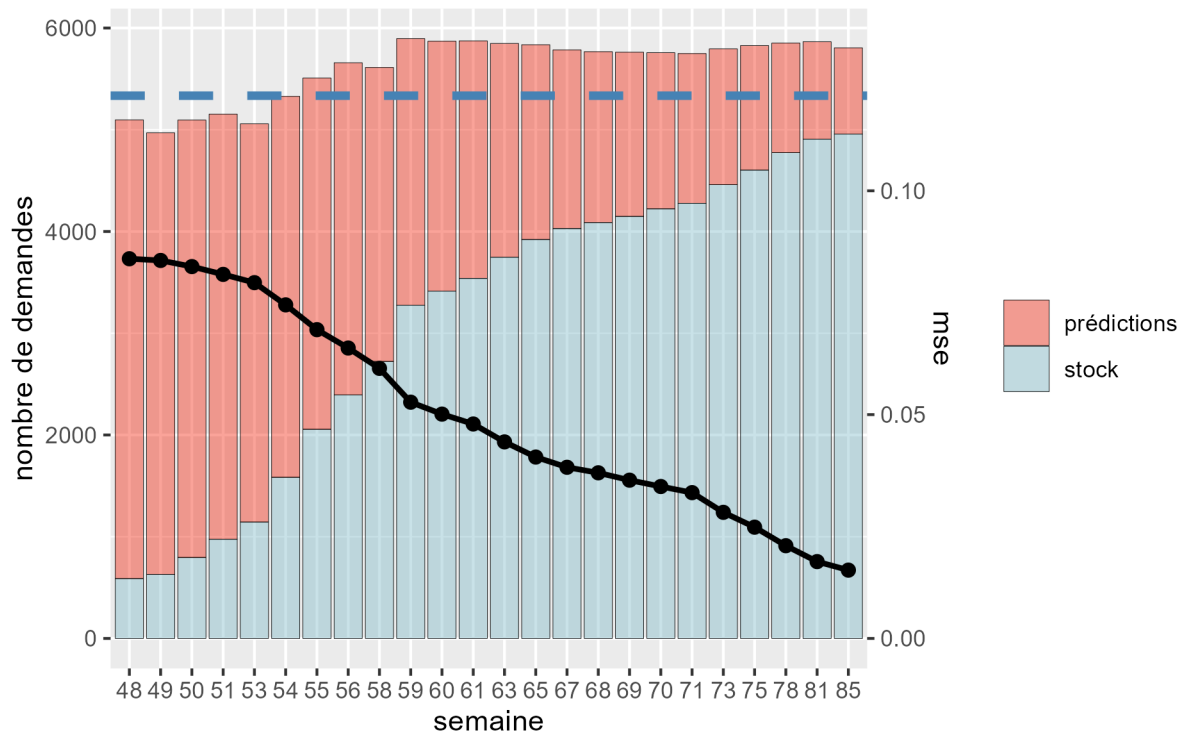


Figure 3.13: *MSE, stock de demandes et somme des prédictions des probabilités de demande de reconnaissance au titre de l'événement sécheresse 2020 pour chaque semaine étudiée. La MSE est représentée par la courbe correspondant à la fonction $u \mapsto \frac{1}{|\mathcal{A}_{2020,u}^-|} \sum_{\alpha \in \mathcal{A}_{2020,u}^-} (\zeta_{\alpha,2020} - \ell_{5,2019}(X_{\alpha,2020,u}))^2$. Le trait bleu en pointillés représente le nombre de demandes réelles. Les prédictions sont réalisées par un meta-algorithme fondé sur un réseau de neurones profond ($k = 5$).*

Chapter 4

Making sparse predictions, and forecasting the requests of the government declaration of natural disaster for a drought event in France

Nous proposons dans ce chapitre une méthode alternative à l'OSASSL pour l'anticipation des communes demanderesses de la reconnaissance de l'État de catastrophe naturelle au titre du péril sécheresse. Cette proposition est motivée par le souhait d'adopter un autre point de vue pour la résolution de ce problème. Le transport optimal computationnel s'étant révélé très efficace dans de nombreux cadres applicatifs [voir par exemple la monographie Peyré and Cuturi, 2020], nous avons fait le choix de fonder notre nouvelle démarche sur cette théorie. Par ailleurs, en favorisant la production de prédictions nulles au détriment d'estimations de faibles probabilités, l'algorithme présenté dans ce chapitre tient compte dès sa conception de la rareté des demandes.

Ce chapitre est le fruit d'une collaboration avec Thi Thanh Yen Nguyen (doctorante au laboratoire MAP5) et Antoine Chambaz (notre directeur de thèse). Du fait de ma maîtrise du problème considéré, j'ai joué un rôle déterminant dans sa modélisation, dans la préparation des données, dans la conception de l'algorithme hybride et dans l'analyse des résultats.

Ce chapitre fera prochainement l'objet d'une prépublication. Il sera alors soumis à une revue internationale.

4.1 Introduction

We define a drought event in this study as the phenomenon of clay shrinking and swelling during a calendar year. For a comprehensive introduction to drought events and their economic consequences, we refer to [Charpentier et al., 2022, Sections 1 and 2]. In brief, the clay in the soil undergoes alternating shrinkage and swelling in dry and humid conditions, leading to instabilities and cracks in buildings. The costs incurred by these cracks are covered by private property insurance policies. As 90% of the French natural disasters insurance market is reinsured by [Caisse Centrale de Réassurance](#) (henceforth abbreviated as CCR) [CCR, 2022b], a public-sector reinsurer providing coverage against natural catastrophes and uninsurable risks, the French state ultimately bears part of the risk.

Due to intricacies of the French legal framework [known as the natural disasters compensation scheme, see Charpentier et al., 2022, Section 2.1], two prerequisites must be met in order to initiate the compensation scheme. Firstly, the property that has been lost and/or damaged must be covered by a property and casualty insurance policy, which is a condition of private nature. Secondly, a government decree declaring a natural disaster must be published in the Official Journal, which is a condition of public nature. The responsibility of initiating the request for the government declaration of a natural disaster for the cities they administer lies with the mayors. Of note, we adopt here and henceforth the term “city” regardless of the size of the *commune*, encompassing a wide range from small hamlets to large urban centers.

Forecasting the cost of drought events in France is a critical task for CCR. CCR currently addresses two sub-problems separately: sub-problem 1 involves predicting which cities will submit a request for the government declaration of natural disaster for a drought event, while sub-problem 2 is centered on predicting the cost of a drought event for those cities that have already obtained the government declaration of natural disaster for a drought event. In this study, we concentrate on sub-problem 1. [Ecoto et al., 2021a, Ecoto and Chambaz, 2022a] focus on sub-problem 2. In contrast, [Chatelain and Loisel, 2021] takes on both sub-problems simultaneously. On the other hand, [Charpentier et al., 2022, Heranval et al., 2022] predict which cities will experience claims (a proxy for sub-problem 1) and subsequently estimate the cost for these cities. We acknowledge that the problem we address in this study is, therefore, more narrowly focused than those studied in [Chatelain and Loisel, 2021, Charpentier et al., 2022, Heranval et al., 2022].

Quoting [Logar and van den Bergh, 2011, page 4, first paragraph], “[t]he existing literature on the costs of drought [events] is scarce, fragmented and heterogeneous and there is a need for comprehensive costs estimations to help designing effective policy responses.” To the best of our knowledge, [Chatelain and Loisel, 2021, Charpentier et al., 2022, Heranval et al., 2022, Ecoto et al., 2021a, Ecoto and Chambaz, 2022a] are the only five references available about the prediction of the cost of drought events, thus susceptible to address the problem of predicting which cities will submit a request for the government declaration of natural disaster for a drought event. It is worth noting that studies conducted by insurance companies are often kept confidential, further emphasizing the

scarcity of available literature on this subject.

In [Chatelain and Loisel, 2021], the authors use Generalized Linear Models (GLM) and the extreme gradient boosting algorithm to predict which cities will submit a request for the government declaration of natural disaster for a drought event (see Section 3.1 therein). We also tackle the problem as a classification task, leveraging the power of classification algorithms. However, taking a slightly different perspective, our main contribution consists in introducing an alternative procedure that hinges on optimal transport theory and an inertial proximal algorithm for nonconvex optimization. The optimization problem is designed so as to yield a sparse vector of predictions because it is known that relatively few cities will submit requests. Additionally, we develop a hybrid procedure that synergistically combines and utilizes both types of predictions.

The rest of the study is organized as follows. Section 4.2 introduces the data set that we obtained by merging several data sets, some of which either provided by CCR’s cedents⁶ while others were collected from other trusted sources. This section also outlines the statistical challenge that we undertake and presents insights into the data. Section 4.3 is a modicum of optimal transport theory. Section 4.4 exposes our novel procedure to make sparse predictions and discusses how to solve the nonconvex optimization task that sits at its core using the algorithm iPiano [Ochs et al., 2015], from both theoretical and computational perspectives. Section 4.5 presents a simulation study and introduces the hybrid procedure. Section 4.6 describes the full-fledged application to the challenge of forecasting which cities will submit a request for the government declaration of natural disaster for a drought event. Section 4.7 discusses our results and outlines potential avenues for future research. In the appendix, Section 4.8 gathers the proofs of the convergence of the iPiano algorithm using a theorem proven in [Ochs et al., 2015]. The Kurdyka-Lojasiewicz property [Attouch et al., 2010] and notion of ϵ -minimal structures [Wilkie, 1996] play a central role.

4.2 Data and statistical challenge

4.2.1 Presentation of the data, first pass

The data set is obtained by merging several data sets, either provided by CCR’s cedents or collected from other sources, namely the National Institute for Statistical and Economic Studies (Insee), Geographic National Institute (IGN), French Geological Survey (BRGM) and Météo-France. While there are numerous similarities between the present data set and the one comprehensively presented and used in [Ecoto and Chambaz, 2022a, see Section 2], there are also major differences.

From now on, France refers to *Metropolitan* or *Mainland* France, and the adjective French to what is related to France with the restricted acceptation of the word. This is justified because drought events are not a threat in Overseas France (essentially because there is

6. A cedent is a party in an reinsurance contract that passes the financial obligation for certain potential losses to the reinsurer. In return for bearing a particular risk of loss, the cedent pays a reinsurance premium.

little clay in these parts of the country).

The experimental units are the French cities. Each of them can contribute a data structure for a given year t (by convention, $t = 1, 2, 3$ respectively correspond to years 2019, 2020 and 2021) and a given week u (the integer $u \in \mathcal{U}_t \subset \mathbb{N}^*$ being the number of weeks starting from the first week of year t , with $44 \leq u \leq 85$). A data structure encompasses multiple aspects of a city's profile, aiming to provide a comprehensive representation of its context and potential triggers for requesting the government declaration of natural disaster for a drought event. It consists of the following blocks of variables:

City description (16 variables). This block provides detailed information about the city, covering various aspects such as housing stock age, housing stock exposure to clay-shrinkage-swelling hazard, and climatic zone. By capturing these variables, a holistic understanding of the city's characteristics is obtained.

City exposure to drought events (25 variables). The variables within this block outline the city's exposure to drought events. They build upon the Soil Wetness Index (SWI), and include an indicator of whether or not the city is eligible for the government declaration of natural disaster for a drought event.

City history of requests (12 variables). This block provides a record of the city's previous requests for the government declaration of natural disaster for a drought event, including information on the success or failure of the requests. The record gives us insight into the city's decision-making process, intentions and actions regarding the submission of a request for the government declaration of natural disaster for a drought event.

City current request status (1 variable). This variable indicates whether or not the city submitted a request for the government declaration of natural disaster for a drought event for year t during week u or before.

City's vicinity description (13 variables). This block focuses on the city's surroundings. It provides information about the neighboring cities' claims and requests for the government declaration of natural disaster for a drought event.

4.2.2 Presentation of the data, second pass

Description of a city. The description of a city notably consists of its population, of the (estimated) number of houses located within the city's limits [the estimation is based on census data: Insee, 2000], of the city's average altitude and area [source: IGN, 2018], house density (defined as the ratio of the number of houses to the city's area), and proportions of buildings built prior to 1949, between 1950 and 1974, between 1975 and 1989, and after 1989 [the proportions are computed based on data found in Insee, 2000]. In addition, the description of the city also includes the proportions of houses located within the city's limits that fall in each of the four clay-shrinkage-swelling hazard categories [as defined by, and obtained from BRGM: MI, 2019]; the city's seismic zone (a four-category variable attributed to each city by the French *Code de l'environnement*); the climatic zone of the city's department (the French State attributes

to each department this five-category variable; a department is a level of government between the administrative regions and communes).

Up to now, the variables that we listed are essentially static. The description of the city is completed by the (estimated) insured sum corresponding to the houses located within its limits. The estimations are based on data from Insee and portfolios data provided by CCR's cedents. This last piece of information depends on the year, but the variations from one year to another are limited.

To conclude, let us stress that the age of the housing stock is used here as a proxy for the house building technology, an important factor to consider because some buildings are more vulnerable than others [France Assureurs, 2022, page 28]. Furthermore, accounting for clay concentration is mandatory since it is the clay present in the soil that, by shrinking and swelling in dry and humid conditions, creates instabilities and generates cracks in buildings.

Description of a city's exposure to a specific drought event. The description of a city's exposure to a specific drought event builds upon the SWI in a manner presented almost comprehensively in [Ecoto and Chambaz, 2022a, Section 2.3.2]. For self-containedness, we recall here the main elements of the presentation.

Provided by Météo-France since 1959, the SWI data consist of time series of values (one value every ten-day period) ranging between -3.33 (very dry soil) and 2.33 (very wet soil). There are as many SWI time series as the number of 8×8 km² squares used by Météo-France to partition the French territory.

Note that for any year t and week $u \in \mathcal{U}_t \cap \llbracket 44, 52 \rrbracket$ (that is, before the end of year t), we necessarily have access to fewer than 36 values of the SWI for year t . We use a prediction model [Ardon, 2014] to predict future values of the SWI so that all the time series of SWI cover the whole year. As u increases, the predicted values are replaced by the actual values provided by Météo-France, until the complete time series for year t are all observed.

For every year t and every city, we then derive a city-specific SWI time series by taking the convex average of the possibly completed SWI time series attached to the squares that overlap the city's area, the weights being proportional to the areas of the intersections. The description of a city's exposure to drought events for year t builds upon the corresponding SWI time series. It notably consists of the minimum value of the SWI time series, of the overall average of the time series, of the averages restricted to the first, second, third and fourth quarters of year t respectively (that is, January-March, April-June, July-September, October-December), and of the averages restricted to the unions of the second and third quarters (April-September) or of the first, second and third quarters (January-September). The description is complemented by measures of how exceptional the monthly and quarterly average SWI (say $\overline{\text{SWI}}$) are relative to historical SWI data. Specifically, for every month (respectively, every quarter), we compute the empirical cumulative distribution function of the monthly (respectively, quarterly) average SWI using all data for the city of interest from 1959 to 2009 and then evaluate that function at $\overline{\text{SWI}}$. The smaller is the resulting proportion, the more pronounced is

the soil dryness and, conversely, the larger is the resulting proportion, the more pronounced is the soil wetness. Moreover, the description includes an indicator of whether or not the city is eligible for a government declaration of natural disaster for a drought event.

This description holds utmost relevance as it focuses on the critical role of soil humidity in causing the shrinkage and swelling of clay, eventually leading to instabilities and the formation of cracks in buildings.

Requests for the government declaration of natural disaster for a drought event. Being the secretary of the Commission Interministérielle Catastrophe Naturelle, CCR has been having access, since 1989, to the requests for the government declaration of natural disaster for a drought event as they accrue. Formally, a city can submit a request for the government declaration of natural disaster for a drought event for year t until the end of June of year $(t + 2)$. However, anticipating which cities will submit a request for year t is only a necessity typically between the months of November of year t and of December of year $(t + 1)$.

Description of a city's request history. Given a year t and a week u , the (t, u) -specific description of a city's request history consists of t and u , of the overall number of French cities that submitted a request for year t during week u or before, and of the ratio of the logarithm of that overall number to u . In addition, the description includes the number of requests submitted by the city since 1990 (respectively, between years $(t - 4)$ and t), the number of times the city obtained the government declaration of natural disaster for a drought event since 1990 (respectively, between years $(t - 4)$ and t), and the ratio of the aforementioned number of requests submitted by the city since 1990 to the number of years between 1990 and year t . Moreover, the description includes an indicator of whether or not the city was denied the government declaration of natural disaster for a drought event on year $(t - 1)$, and the numbers of denied requests between $(t - 2)$ and $(t - 1)$ and between $(t - 4)$ and $(t - 1)$.

This description holds significant relevance, primarily due to its ability to provide valuable insights into the city's inclination to submit a request for a government declaration of natural disaster for a drought event. By examining the city's historical pattern of submitting such requests since 1990 or within the previous five years, regardless of their success, we can gather essential information about the city's familiarity with the administrative procedure. Additionally, this serves as a proxy for assessing the city's exposure to drought events.

Description of a city's vicinity. Using the flux of requests, we compile a collection of variables describing the vicinity of a city. The variables concern either the neighboring cities or, more broadly, the cities in the same department. Given a year t and a week u , the (t, u) -specific collection notably consists of the following five numbers: the number of neighboring cities that requested the government declaration of natural disaster for a drought event for year t during week u or before, the number of neighboring cities

(respectively, of cities in the same department) that submitted such a request *for the first time* for year t , and the number of neighboring cities (respectively, of cities in the same department) that submitted such a request *for the first time* between years $(t - 4)$ and t . The collection is complemented by the ratios of the four last numbers to either the number of neighboring cities or the number of cities in the same department. In addition, the collection also includes the number of claims for year t made during week u or before by the neighboring cities (respectively, by the cities of the same department), and the ratio of that number to the number of neighboring cities (respectively, of cities in the same department).

To conclude, it is important to emphasize the potential relevance of these variables for several compelling reasons. For instance, it is common for mayors of neighboring cities to exchange information, particularly if their cities are part of the same federation of municipalities. This interconnectedness means that if a city submits a request for a government declaration of natural disaster for a drought event, then that raises the likelihood that neighboring cities will do the same, either in the same year or later. Furthermore, it is worth noting that drought events are not necessarily confined to a single city's territory. Even if the mayors do not actively share information, the occurrence of a drought event in one city that prompts the submission of a request for a government declaration of natural disaster for a drought event increases the likelihood that a similar drought event has taken place in nearby areas. Consequently, the likelihood of submitting a request for such a declaration also increases in those affected vicinity areas.

4.2.3 The statistical challenge and some facts about the data

As elaborated in Section 4.2.1, each French city can contribute a data structure for a given year t and a given week u (the integer u being the number of weeks starting from the first week of year t). It is worth mentioning that the composition of the set of French cities undergoes slight changes from one year to another. To address this variability, we define \mathcal{A}_t as the set of cities for year t (with the aforementioned convention $t = 1, 2, 3$ for years 2019, 2020 and 2021, respectively). Furthermore, we introduce \mathcal{U}_t as the comprehensive list of weeks during which CCR received the latest submissions of a request for the government declaration of natural disaster for a drought event for year t , encompassing a period of up to 85 weeks following the first week of year t .

We report that $\text{card } \mathcal{A}_1 = \text{card } \mathcal{A}_2 = 34,841$ and $\text{card } \mathcal{A}_3 = 34,836$. Moreover,

$$\mathcal{U}_1 = \{44, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 58, 59, 60, 61, 69, 75\},$$

$$\mathcal{U}_2 = \{48, 49, 50, 51, 53, 54, 55, 56, 58, 59, 60, 61, 63, 65, 67, 68, 69, 70, 71, 73, 75, 78, 81, 85\},$$

$$\mathcal{U}_3 = \{49, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 64, 65, 66, 67, 68, 71, 72, 73, 77, 78\}.$$

For every year $t = 1, 2, 3$ and each week $u \in \mathcal{U}_t$, we let

- $\xi_{\alpha,t,u} \in \mathcal{X} \subset \mathbb{R}^d$ be city α 's vector of covariates on week u relative to year t (for any city $\alpha \in \mathcal{A}_t$);

- $\zeta_{\alpha,t,u} \in \{0,1\}$ be the indicator equal to 1 if and only if (iff) city α submitted a request *before or during* week u relative to year t (for any city $\alpha \in \mathcal{A}_t$);
- $u^- := \max\{\nu \in \mathcal{U}_t : \nu < u\}$ index the week before u in \mathcal{U}_t (with convention $u^- = 0$ if $u = \min \mathcal{U}_t$), so that $(\zeta_{\alpha,t,u} - \zeta_{\alpha,t,u^-}) \in \{0,1\}$ equals 1 iff city α submitted a request during week u relative to year t (for any city $\alpha \in \mathcal{A}_t$, with convention $\zeta_{\alpha,t,0} = 0$).

In addition we also define, for each year $t = 1, 2, 3$ and any city $\alpha \in \mathcal{A}_t$, $\zeta_{\alpha,t} \in \{0,1\}$, the indicator equal to 1 iff city α submitted a request relative to year t (possibly after the week $\max \mathcal{U}_t$). Note that $\zeta_{\alpha,t} \geq \max_{u \in \mathcal{U}_t} \zeta_{\alpha,t,u}$. In words, some cities may submit a request for the government declaration of natural disaster for a drought event relative to year t beyond week $\max \mathcal{U}_t$. This fact is discussed further in the next paragraph.

Table 4.1 reports the quartiles of the sets

$$\left\{ \sum_{\alpha \in \mathcal{A}_t} (\zeta_{\alpha,t,u} - \zeta_{\alpha,t,u^-}) : u \in \mathcal{U}_t \right\}, \quad t = 1, 2, 3,$$

that is, the quartiles of the sets of the week-specific numbers of new requests for the government declaration of natural disaster for a drought event relative to year t , for $t = 1, 2, 3$. Table 4.1 also reports the initial numbers and proportions of requests for the government declaration of natural disaster for a drought event relative to year t (that is, $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t,\min \mathcal{U}_t}$ and $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t,\min \mathcal{U}_t} / \text{card } \mathcal{A}_t$), their overall numbers and proportions at week $\max \mathcal{U}_t$ (that is, $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t,\max \mathcal{U}_t}$ and $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t,\max \mathcal{U}_t} / \text{card } \mathcal{A}_t$), and the overall numbers and proportions of requests for the government declaration of natural disaster for a drought event relative to year t (that is, $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t}$ and $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t} / \text{card } \mathcal{A}_t$), for $t = 1, 2, 3$. We emphasize that only 12.5% (776/6240), 11.0% (589/5335) and 4.8% (81/1696) of the requests for the government declaration of natural disaster for a drought event relative to year t were already submitted at week $\min \mathcal{U}_t$, while only 82% (5142/6240), 92.9% (4958/5335) and 69.0% (1169/1696) of the overall numbers of requests for the government declaration of natural disaster for a drought event relative to year t were submitted at week $\max \mathcal{U}_t$, for $t = 1, 2, 3$. Moreover, between the first and last weeks $\min \mathcal{U}_t$ and $\max \mathcal{U}_t$, the median numbers of newly submitted requests corresponded to 4.7% (245/5142), 3.3% (166/4958) and 4% (47/1169) of the overall numbers of requests at week $\max \mathcal{U}_t$, for $t = 1, 2, 3$.

Our ultimate objective is to achieve sequential forecasting of which cities will submit a request for the government declaration of natural disaster for a drought event leveraging past data and, in particular, knowing which cities already did. Formally, our objective is the following: for every $u \in \mathcal{U}_3$, leveraging past observations, that is

$$\{(\xi_{\alpha,t,\nu}, \zeta_{\alpha,t,\nu}, \zeta_{\alpha,t}) : t = 1, 2, \alpha \in \mathcal{A}_t, \nu \in \mathcal{U}_t \text{ st } \zeta_{\alpha,t,\nu} = 0\}$$

if $u = \min \mathcal{U}_3$ and otherwise

$$\begin{aligned} & \{(\xi_{\alpha,t,\nu}, \zeta_{\alpha,t,\nu}, \zeta_{\alpha,t}) : t = 1, 2, \alpha \in \mathcal{A}_t, \nu \in \mathcal{U}_t \text{ st } \zeta_{\alpha,t,\nu} = 0\} \\ & \cup \{(\xi_{\alpha,3,\nu}, \zeta_{\alpha,3,\nu}, 0) : \alpha \in \mathcal{A}_3, \nu \in \mathcal{U}_3, \nu < u \text{ st } \zeta_{\alpha,3,\nu} = 0\}, \end{aligned} \quad (4.1)$$

numbers of new requests ($\sum_{\alpha \in \mathcal{A}_t} (\zeta_{\alpha,t,u} - \zeta_{\alpha,t,u^-}), u \in \mathcal{U}_t$)	2019 ($t = 1$)	2020 ($t = 2$)	2021 ($t = 3$)
minimum	104	41	10
1st quartile	138	75	32
median	245	166	47
3rd quartile	386	208	69
maximum	776	589	129
initial number (and proportion) of requests ($\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t,\min \mathcal{U}_t}$)	776 (2.2%)	589 (1.7%)	81 (0.2%)
overall number (and proportion) of requests ($\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t,\max \mathcal{U}_t}$)	5142 (14.8%)	4958 (14.2%)	1169 (3.3%)
overall number (and proportion) of requests ($\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t}$)	6240 (17.9%)	5335 (15.3%)	1696 (4.9%)

Table 4.1: Summary measures of the sets $\{\sum_{\alpha \in \mathcal{A}_t} (\zeta_{\alpha,t,u} - \zeta_{\alpha,t,u^-}) : u \in \mathcal{U}_t\}$ ($t = 1, 2, 3$), that is, of the numbers of new requests for the government declaration of natural disaster for a drought event as weeks go by, for years 2019, 2020 and 2021 respectively. In addition, the overall numbers $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t}$ and proportions $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,t} / \text{card } \mathcal{A}_t$ ($t = 1, 2, 3$) of requests for the government declaration of natural disaster for a drought event relative to year t are also reported for years 2019, 2020 and 2021.

we wish to predict $\zeta_{\alpha,3}$ using $\xi_{\alpha,3,u}$ for every $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. Of note, the set defined in (4.1) when $u = \max \mathcal{U}_3$ consists of more than 2.05 million triplets.

The focus on “making sparse predictions” which is explicit in the title of the manuscript is justified by the last row of Table 4.1: in 2019, 2020 and 2021, the proportions of cities that eventually submitted a request for the government declaration of natural disaster for a drought event were respectively 17.9%, 15.3% and 4.9%.

4.3 A modicum of optimal transport theory

This section introduces the few tools from optimal transport theory that will be instrumental in developing our novel procedure in the next section.

Fix arbitrarily two integers $R, R' \geq 2$. Let $\mathbf{z} := (z_1, \dots, z_R)$ and $\mathbf{z}' := (z'_1, \dots, z'_{R'})$ be two collections of elements of a space \mathcal{Z} . Let $c : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}_+$ map any couple (z, z') to a nonnegative number interpreted as the cost to move z to z' , a cost function. The cost function c induces the $R \times R'$ matrix $C(\mathbf{z}, \mathbf{z}') \in \mathbb{R}_+^{R \times R'}$ whose (r, r') -specific component $(C(\mathbf{z}, \mathbf{z}'))_{r,r'} := c(z_r, z'_{r'})$ is interpreted as the cost to move z_r to $z'_{r'}$ (relative to c).

Let $\Pi_{R,R'} := \{P \in \mathbb{R}_+^{R \times R'} : P \mathbf{1}_{R'} = \frac{1}{R} \mathbf{1}_R, P^\top \mathbf{1}_R = \frac{1}{R'} \mathbf{1}_{R'}\}$ represent the joint laws on $\llbracket R \rrbracket \times \llbracket R' \rrbracket$ with uniform marginal laws, where $\llbracket d \rrbracket := \{1, \dots, d\}$ for every integer $d \geq 1$. For each $P \in \Pi_{R,R'}$, let

$$E(P) := - \sum_{r \in \llbracket R \rrbracket, r' \in \llbracket R' \rrbracket} P_{r,r'} \log P_{r,r'}$$

denote the entropy of P . For every $P \in \Pi_{R,R'}$ and $C \in \mathbb{R}_+^{R \times R'}$, let

$$\langle P, C \rangle := \sum_{r \in \llbracket R \rrbracket, r' \in \llbracket R' \rrbracket} P_{r,r'} \times C_{r,r'}.$$

When $C = C(\mathbf{z}, \mathbf{z}')$, $\langle P, C \rangle$ is interpreted as the (P, C) -specific cost to transport \mathbf{z} onto \mathbf{z}' .

For any $\gamma > 0$ and $C \in \mathbb{R}_+^{R \times R'}$, introduce

$$\mathcal{W}_\gamma(C) := \min_{P \in \Pi_{R,R'}} [\langle P, C \rangle - \gamma E(P)]. \quad (4.2)$$

In particular, when $C = C(\mathbf{z}, \mathbf{z}')$, $\mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z}'))$ is the γ -regularized optimal cost to transport \mathbf{z} onto \mathbf{z}' , abbreviated to “the γ -regularized OT cost”. Considering the γ -regularized OT cost $\mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z}'))$ instead of the regular OT cost $\mathcal{W}_0(C(\mathbf{z}, \mathbf{z}'))$ (defined as in (4.2) with $\gamma = 0$) has two important merits [Peyré and Cuturi, 2020, Chapters 3, 4, 9]. First, $\mathbb{R}_+^{R \times R'} \ni C \mapsto \mathcal{W}_0(C) \in \mathbb{R}$ is not differentiable whereas $\mathbb{R}_+^{R \times R'} \ni C \mapsto \mathcal{W}_\gamma(C) \in \mathbb{R}$ is differentiable. Second, for any $C \in \mathbb{R}_+^{R \times R'}$, computing $\mathcal{W}_0(C)$ requires solving a costly linear program via network simplex methods whereas computing $\mathcal{W}_\gamma(C)$ can be performed easily thanks to the so-called Sinkhorn algorithm [Cuturi, 2013].

Finally, we use the γ -regularized OT cost to define the γ -regularized Sinkhorn cost

$$\mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}') := \mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z}')) - \frac{1}{2} [\mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z})) + \mathcal{W}_\gamma(C(\mathbf{z}', \mathbf{z}'))]$$

(the dependence of $\mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}')$ on the cost function c is hidden). By [Feydy et al., 2019b, Theorem 1], $\mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}') \geq \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}) = 0$. Moreover, we stress that $\mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}')$ can be computed with little additional computational cost compared to $\mathcal{W}_\gamma(\mathbf{z}, \mathbf{z}')$.

4.4 Making sparse predictions

The procedure we are about to present is funded on two core ideas. Firstly, we aim to predict whether a city will submit a request for the government declaration of natural disaster for a drought event by employing an interpretable comparison of the city’s covariates with those of other cities whose submission status may be already known. Secondly, we want to have a control on the sparsity of the set of predictions and encourage 0-predictions, which correspond to cases where we predict that a city will not submit a request.

4.4.1 Translation to an optimization problem

As elaborated in Section 4.2.3, our objective is to predict $\zeta_{\alpha,3}$ based on $\xi_{\alpha,3,u}$ for every $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$, using past observations (4.1), and so repeatedly for each $u \in \mathcal{U}_3$. In the rest of the study, it will be convenient to denote generically $\{(x_m, y_m) : m \in \llbracket M \rrbracket\} \subset \mathcal{X} \times \{0, 1\}$ and $\{(x'_n, y'_n) : n \in \llbracket N \rrbracket\} \subset \mathcal{X} \times \{0, 1\}$ two collections of couples for which it is desired to predict y'_n based on x'_n , for every $n \in \llbracket N \rrbracket$, using

past observations $(x_1, y_1), \dots, (x_M, y_M)$. To do so, we propose to solve the following optimization problem:

$$\arg \min_{\theta \in \mathbb{R}^N} \{ \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}'(\theta)) + g_\tau(\theta) \}, \quad (4.3)$$

where

— for all $\theta \in \mathbb{R}^N$,

$$\mathbf{z} := ((x_1, y_1), \dots, (x_M, y_M)), \quad \mathbf{z}'(\theta) := ((x'_1, \theta_1), \dots, (x'_N, \theta_N));$$

— the cost function $c : (\mathcal{X} \times \mathbb{R}) \times (\mathcal{X} \times \mathbb{R}) \rightarrow \mathbb{R}_+$ is given by

$$c((x, y), (x', \theta)) := \text{dis}(x, x')^2 + (y - \theta)^2 \quad (4.4)$$

for a distance or dissimilarity dis on \mathcal{X} ;

— g_τ is a convex function given by either $g_\tau(\theta) := \tau \|\theta\|_1 + \mathbf{I}\{\theta \in [0, 1]^N\}$, with $\|\theta\|_1 := \sum_{n \in [N]} |\theta_n|$, or $g_\tau(\theta) := \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0, 1]^N\}$, where $\mathbf{I}\{A\}$ equals 0 if A is true and $+\infty$ otherwise;

— $\gamma, \tau > 0$ are some user-supplied constants.

A few comments are in order. Firstly, the argmin in (4.3) is over \mathbb{R}^N but could equivalently be over $[0, 1]^N$ (even if the term $\mathbf{I}\{\theta \in [0, 1]^N\}$ was dropped from the definitions of $g_\tau(\theta)$). We thus view θ_n as the probability that the city described by x'_n will submit a request of the government declaration of natural disaster for a drought event.

Secondly, though hidden in the notation, the cost function c obviously plays a pivotal role. It operationalizes the core idea of making predictions based on comparisons between the covariates of different cities.

Thirdly, for both choices of g_τ , the ℓ^1 -norm of θ can be seen as a measure of sparsity of θ , a substitute for the integer $\text{card}\{n \in [N] : \theta_n \neq 0\}$. Incorporating the penalization term $+g_\tau(\theta)$ operationalizes the core idea of promoting sparse solutions, aligning with our prior understanding that only a limited number of cities will eventually submit a request of the government declaration of natural disaster for a drought event (see Table 4.1 for the actual numbers and proportions of cities that did in 2019, 2020 and 2021). Finally, the case where $g_\tau(\theta) = \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0, 1]^N\}$ is quite interesting because, as we will see, there is a natural way to select τ .

4.4.2 On solving (4.3)

Solving (4.3) is not straightforward, in part because the criterion to minimize is the sum of the non-convex differentiable function $f : \theta \mapsto \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}'(\theta))$ (see Section 4.8.1.2) and of the convex non-differentiable function g_τ . Luckily, we can rely on the so-called iPiano algorithm [Ochs et al., 2015] which was developed precisely to deal with such optimization problems.

The iPiano algorithm starts from an initial $\theta^{-1} = \theta^0 \in]0, 1[^N$ and the update scheme informally writes as (below, α, β are positive constants)

$$\theta^{k+1} = \text{Prox}_{\alpha g_\tau} \left(\theta^k - \alpha \nabla f(\theta^k) + \beta(\theta^k - \theta^{k-1}) \right), \quad (4.5)$$

where the proximal map $\text{Prox}_{\alpha g_\tau}$ is defined by

$$\text{Prox}_{\alpha g_\tau}(t) := \arg \min_{\theta \in \mathbb{R}^N} \left\{ \frac{1}{2} \|\theta - t\|_2^2 + \alpha g_\tau(\theta) \right\}. \quad (4.6)$$

On the one hand, if $g_\tau(\theta) = \tau \|\theta\|_1 + \mathbf{I}\{\theta \in [0, 1]^N\}$ then (4.6) is simply given by

$$(\text{Prox}_{\alpha g_\tau}(t))_n = \min\{(|t_n| - \alpha\tau)_+, 1\}.$$

In particular, if $t \in [0, 1]^N$ then $(\text{Prox}_{\alpha g_\tau}(t))_n = (t_n - \alpha\tau)_+$ for every $n \in \llbracket N \rrbracket$. On the other hand, if $g_\tau(\theta) = \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0, 1]^N\}$ then the proximal map is the Euclidean projection onto the ℓ^1 -ball centered at 0 and with radius τ . An efficient algorithm is available to implement this projection [Duchi et al., 2008].

Moreover, following [Cuturi and Doucet, 2014, Section 4.3], we show in Section 4.8.1.2 that the gradient of f is given by

$$\begin{aligned} \nabla f(\theta) &= \nabla \mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z}'(\theta))) - \frac{1}{2} \nabla \mathcal{W}_\gamma(C(\mathbf{z}'(\theta), \mathbf{z}'(\theta))) \\ &= 2\left(\frac{1}{N}\theta - \widehat{P}_\theta^\top y\right) - \left(\frac{2}{N}\theta - (\widehat{Q}_\theta + \widehat{Q}_\theta^\top)\theta\right) \\ &= -2\widehat{P}_\theta^\top y + (\widehat{Q}_\theta + \widehat{Q}_\theta^\top)\theta \end{aligned} \quad (4.7)$$

with

$$\widehat{P}_\theta = \arg \min_{P \in \Pi_{M,N}} \{ \langle P, C(\mathbf{z}, \mathbf{z}'(\theta)) \rangle - \gamma E(P) \}, \quad (4.8)$$

$$\widehat{Q}_\theta = \arg \min_{P \in \Pi_{N,N}} \{ \langle P, C(\mathbf{z}'(\theta), \mathbf{z}'(\theta)) \rangle - \gamma E(P) \}. \quad (4.9)$$

We check that the assumptions of [Ochs et al., 2015, Theorems 4.9 and 4.14] are met by proving that f is C^1 -smooth with an L -Lipschitz gradient on $\text{dom } g_\tau$ and that $(f + g_\tau)$ satisfies the Kurdyka-Lojasiewicz property on its domain (the proof is presented in Section 4.8). Therefore we can assert that

- the sequence $(\theta^k)_{k \geq 0}$ converges to a critical point of $\theta \mapsto f(\theta) + g_\tau(\theta)$;
- $\min_{k \leq K} \|\theta^{k+1} - \theta^k\|_2^2 = O(K^{-1})$;
- if we set $r(\theta) := \theta - \text{Prox}_{\alpha g_\tau}(\theta - \alpha \nabla f(\theta))$, then $\min_{k \leq K} \|r(\theta^k)\|_2^2 = O(K^{-1})$.

The so-called proximal residual $r(\theta)$ is interesting because $r(\theta) = 0$ means that the first-order optimality condition is met at θ . Indeed (denoting by $\partial \ell(x)$ either the subdifferential of the convex function ℓ at x or the limiting-subdifferential of the proper lower semicontinuous function ℓ at x , see Section 4.8.2.1), $r(\theta) = 0$ iff

$$\begin{aligned} \theta = \text{Prox}_{\alpha g_\tau}(\theta - \alpha \nabla f(\theta)) &\quad \text{iff} \quad 0 \in \partial \left(\frac{1}{2} \|\theta - \alpha \nabla f(\theta) - \cdot\|_2^2 + \alpha g_\tau \right) (\theta) \\ &\quad \text{iff} \quad 0 \in \{ \theta - (\theta - \alpha \nabla f(\theta)) \} + \alpha \partial g_\tau(\theta) \\ &\quad \text{iff} \quad 0 \in \{ \alpha \nabla f(\theta) \} + \alpha \partial g_\tau(\theta) \\ &\quad \text{iff} \quad 0 \in \partial (f + g_\tau)(\theta). \end{aligned}$$

4.4.3 Implementation of the “OT-procedure”

Algorithm 1 solves (4.3) by using the iPiano algorithm and a mini-batch procedure to cope with situations where M and N are large. From now on, running the OT-procedure will mean applying Algorithm 1.

Algorithm 1 A mini-batch version of the inertial proximal algorithm for nonconvex optimization (iPiano) tailored to solve (4.3). For any vector $\theta \in \mathbb{R}^N$ and subset \mathcal{N} of $\llbracket N \rrbracket$, we denote $\theta|_{\mathcal{N}} := (\theta_n)_{n \in \mathcal{N}}$.

Input: Data $\{(x_m, y_m) : m \in \llbracket M \rrbracket\}$, $\{x'_n : n \in \llbracket N \rrbracket\}$; regularization parameter $\gamma > 0$, constraint $\tau > 0$; learning rate $\alpha > 0$, momentum parameter $\beta \geq 0$; batch size $B \in \mathbb{N}^*$, number of iterations $T \in \mathbb{N}^*$

Output: Proposed optimizer θ^T

Sample $\theta^{-1} \in \mathbb{R}^N$ with independent components drawn from the uniform law on $[0, 0.01]$

Set $\theta^{-1} \leftarrow 0.5 + \theta^{-1}$ and $\theta^0 \leftarrow \theta^{-1}$

Set $t \leftarrow 0$

while $t < T$ **do**

Independently, sample uniformly without replacement $\mathcal{M} \subset \llbracket M \rrbracket$, $\mathcal{N} \subset \llbracket N \rrbracket$ of cardinality B

Set $\mathbf{z} \leftarrow ((x_m, y_m) : m \in \mathcal{M})$ and $\mathbf{z}'(\theta^t|_{\mathcal{N}}) \leftarrow ((x'_n, \theta_n^t) : n \in \mathcal{N})$

Compute $F(\theta^t|_{\mathcal{N}}) = \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}'(\theta^t|_{\mathcal{N}}))$ using Sinkhorn’s algorithm


Compute $\nabla F(\theta^t|_{\mathcal{N}})$ using automatic differentiation

Set $\theta^{t+1} \leftarrow \theta^t$ and update $\theta^{t+1}|_{\mathcal{N}} \leftarrow \theta^{t+1}|_{\mathcal{N}} - \alpha \nabla F(\theta^t|_{\mathcal{N}}) + \beta(\theta^t|_{\mathcal{N}} - \theta^{t-1}|_{\mathcal{N}})$

Update $\theta^{t+1} \leftarrow \text{Prox}_{\alpha g_\tau}(\theta^{t+1})$

Update $t \leftarrow t + 1$

end while

We wrote a `python/pytorch` program that implements Algorithm 1. Available at , the program hinges on the `GeomLoss` package [Feydy et al., 2019a] which provides a very fast GPU implementation of the Sinkhorn algorithm [Cuturi, 2013].

In Section 4.5, we conduct a simple simulation study in a simple context where $\mathcal{X} = \mathbb{R}^2$ and both M and N are relatively small. We compare the results obtained by aggregating the predictions acquired from classification algorithms with those achieved through the OT-procedure. Notably, we report how we select the pivotal cost function (4.4), g_τ and the hyperparameters (γ, α, β) of Algorithm 1. Moreover, we also introduce the hybrid procedure which synergistically combines and utilizes the two types of predictions.

Section 4.6 is dedicated to the challenging task of forecasting the requests of the government declaration of natural disaster for a drought event. This real-world application poses greater challenges than the simulation study. Tangibly, these challenges arise because $\mathcal{X} \subset \mathbb{R}^d$ is a relatively high-dimensional space ($d = 67$) and both M and N are large. Intangibly, the intricacies lie in the mechanisms that determine whether a request is submitted or not.

We compare the results obtained from a classification algorithm with those achieved

through the OT-procedure and the hybrid procedure. Regarding the OT-procedure, we notably rely on HYPERBAND [Li et al., 2018], a bandit-based approach to hyperparameter optimization, to define the pivotal cost function, and on a simple grid search to then fine-tune the hyperparameters (γ, α, β) of Algorithm 1.

4.5 A simple simulation study, introducing the “hybrid procedure”

4.5.1 Simulated data

For any $p \in (0, 1)$, let P_p be the law on $\mathbb{R}^2 \times \{0, 1\}$ such that

- if R and A are independently drawn from the $\chi^2(1)$ law and from the uniform law on $[0, 2\pi]$, if $X = (R \cos(A), R \sin(A))$ and if, conditionally on X , Y is drawn from the Bernoulli law with parameter $\text{expit}(\text{cst}(p) + R)$, then the joint law of (X, Y) is P_p ;
- the above constant $\text{cst}(p) \in \mathbb{R}$ is defined in such a way that $E_{P_p}(Y) = P_p(Y = 1) = p$.

For instance, $\text{cst}(15\%) \approx -3.13$, $\text{cst}(10\%) \approx -3.83$ and $\text{cst}(5\%) \approx -5.00$. Note that, for any $p \in (0, 1)$, under P_p , the further X is from 0 the more likely it is that $Y = 1$.

We generate independently $L = 30$ data sets as follows. For each $\ell \in \llbracket L \rrbracket$, for every $p \in \{15\%, 10\%, 5\%\}$, we independently sample $n = 1000$ independent copies of (X, Y) under P_p . We thus obtain $M = 3n$ couples $(x_{m,\ell}, y_{m,\ell})$. Moreover, we also sample independently $n = 1000$ independent copies of (X, Y) from the law P_p with $p = 5\%$. We thus obtain $N = n$ couples $(x'_{n,\ell}, y'_{n,\ell})$. Our objective is to recover, for each $\ell \in \llbracket L \rrbracket$, the vector $(y'_{n,\ell})_{n \in \llbracket N \rrbracket}$ based on $\{(x_{m,\ell}, y_{m,\ell}) : m \in \llbracket M \rrbracket\}$ and on $(x'_{n,\ell})_{n \in \llbracket N \rrbracket}$.

4.5.2 Fine-tuning the OT-procedure

Let us first describe how we fine-tune the OT-procedure in order to predict $(y'_{n,\ell})_{n \in \llbracket N \rrbracket}$ by solving (4.3) with $(x_m, y_m) = (x_{m,\ell}, y_{m,\ell})$ and $(x'_n, y'_n) = (x'_{n,\ell}, y'_{n,\ell})$ for all $m \in \llbracket M \rrbracket$ and $n \in \llbracket N \rrbracket$, for each $\ell \in \llbracket L \rrbracket$ in turn. On the one hand, we select the cost function $c : (\mathbb{R}^2 \times \{0, 1\}) \times (\mathbb{R}^2 \times \{0, 1\}) \rightarrow \mathbb{R}_+$ (4.4) given by

$$c((x_1, x_2, y), (x'_1, x'_2, y')) := 100 \times \left| \sqrt{x_1^2 + x_2^2} - \sqrt{(x'_1)^2 + (x'_2)^2} \right| + (y - y')^2.$$

Admittedly, this puts us in a favorable position because the true conditional probability of the event $Y = 1$ given X only depends on $\sqrt{X_1^2 + X_2^2}$. On the other hand, we choose the function $g_\tau : \theta \mapsto \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0, 1]^N\}$ for a τ whose choice is explained in Section 4.5.3. Furthermore, in view of Algorithm 1, we set $\gamma = 10^{-3}$, $\alpha = 10^{-3}$, $\beta = 10^{-4}$, $B = 128$ and $T = 2000$.

4.5.3 Alternative, classification-based approaches

As an alternative approach, we also consider training an algorithm using $\{(x_{m,\ell}, y_{m,\ell}) : m \in \llbracket M \rrbracket\}$ in order to learn to classify each $x'_{n,\ell}$ individually ($n \in \llbracket N \rrbracket$), for every $\ell \in \llbracket L \rrbracket$ in turn. Instead of selecting one algorithm, we rely on super learning to learn and train a meta-algorithm that builds upon several algorithms to classify at least as well as (and sometimes better than) all the candidate algorithms [van der Laan et al., 2007, Polley et al., 2021, 2011, and references therein]. We rely on four individual algorithms to learn the conditional probability of the event $Y = 1$ given X : an algorithm that approximates it under the form of a constant function (in X); an algorithm that learns which element of the working model $\{x \mapsto \text{expit}(t_0 + t_1x_1 + t_2x_2) : t \in \mathbb{R}^3\}$ best approximates it (see `stats::glm`); an algorithm that approximates it under the form of a tree, using the covariates X_1 and X_2 (see `rpart::rpart`); an algorithm that approximates it under the form of a random forest, using the covariates X_1 and X_2 (see `ranger::ranger`) – more details are given below.

In addition, we consider a second super learning procedure to learn the conditional probability of the event $Y = 1$ given X by relying on: an algorithm that approximates it under the form of a constant function (in X); an algorithm that learns which element of the working model $\{x \mapsto \text{expit}(t_0 + t_1x_1 + t_2x_2 + t_3\sqrt{x_1^2 + x_2^2}) : t \in \mathbb{R}^4\}$ best approximates it (see `stats::glm`); an algorithm that approximates it under the form of a tree, using the covariates X_1 , X_2 and $\sqrt{X_1^2 + X_2^2} = R$ (see `rpart::rpart`); an algorithm that approximates it under the form of a random forest, using the covariates X_1 , X_2 and R (see `ranger::ranger`). We expect the second super learner to perform better than the first one because it can use the relevant covariate R .

We use the `SuperLearner` R package [R Core Team, 2022, Polley et al., 2021] to implement and train the super learners. For both super learning procedures, we rely on V -fold cross validation with $V = 10$ folds and use the default hyperparameters specified in `SuperLearner::SL.glm`, `SuperLearner::SL.rpart` [Therneau and Atkinson, 2019] and `SuperLearner::SL.ranger` [Wright and Ziegler, 2017].

4.5.4 Results, introducing the “hybrid procedure”

For each $\ell \in \llbracket L \rrbracket$, we train the two super learners and denote by $\hat{y}'_{n,\ell}{}^{\text{SL}_1}$ and $\hat{y}'_{n,\ell}{}^{\text{SL}_2}$ the estimates of the conditional probabilities that $Y = 1$ given $X = x'_{n,\ell}$ that they output for each $n \in \llbracket N \rrbracket$. Next, we set $\tau = \|\hat{y}'_{n,\ell}{}^{\text{SL}_2}\|_1$ for the OT-procedure, run it, and denote by $\hat{y}'_{n,\ell}{}^{\text{OT}}$ the estimates of the conditional probability that $Y = 1$ given $X = x'_{n,\ell}$ for each $n \in \llbracket N \rrbracket$ that it yields.

Before discussing the results, we introduce a fourth procedure that we aptly refer to as the “hybrid procedure” because it builds upon the OT-procedure and the second super learning procedure. Specifically, the hybrid procedure produces estimates of the above conditional probabilities which are merely defined as the geometric means of the estimates output by the second super learner and yielded by the OT-procedure. Hereafter, these estimates are denoted by $\hat{y}'_{n,\ell}{}^{\text{HYB}} := (\hat{y}'_{n,\ell}{}^{\text{SL}_2} \times \hat{y}'_{n,\ell}{}^{\text{OT}})^{1/2}$ for every $n \in \llbracket N \rrbracket$.

Figure 4.1 provides insights into the predictions $\{\hat{y}'_{n,\ell}^\bullet : n \in \llbracket N \rrbracket\}$ where the symbol \bullet stands for $\text{SL}_1, \text{SL}_2, \text{OT}, \text{HYB}$. On the one hand, the empirical cumulative distribution functions (ecdfs) plotted in the left-hand side panel of Figure 4.1 reveal that the predictions $\hat{y}'_{n,\ell}{}^{\text{OT}}$ for $(n, \ell) \in \llbracket N \rrbracket \times \llbracket L \rrbracket$ such that $y_{n,\ell} = 0$ are often (17%) equal to 0 and are generally more concentrated around 0 than the other predictions (the red ecdf dominates the others). In stark contrast, the predictions $\hat{y}'_{n,\ell}{}^{\text{SL}_1}$ and $\hat{y}'_{n,\ell}{}^{\text{SL}_2}$ for the same couples (n, ℓ) are bounded away from 0 (being larger than 1.56% and 1.35%, respectively). On the other hand, the ecdfs plotted in the right-hand side panel of Figure 4.1 reveal that the predictions $\hat{y}'_{n,\ell}{}^{\text{OT}}$ for $(n, \ell) \in \llbracket N \rrbracket \times \llbracket L \rrbracket$ such that $y_{n,\ell} = 1$ can be equal to 0 (2.7%) and are generally smaller than the other predictions (the red ecdf dominates the others again). They also show that the second super learner outperforms the first one in the sense that the maximum gap between their ecdfs is large (a Kolmogorov-Smirnov viewpoint). Furthermore, by conducting a comparison across panels we discern the notable and desirable trend wherein the predictions $\{\hat{y}'_{n,\ell}^\bullet : n \in \llbracket N \rrbracket, \ell \in \llbracket L \rrbracket \text{ st } y'_{n,\ell} = y\}$ exhibit larger values when $y = 1$ as opposed to when $y = 0$. In conclusion, the hybrid predictions seem to strike a fine balance between the predictions output by the second super learner and the OT-procedure.

In order to complement this first analysis, we employ mean squared error (MSE) as a measure of performance and compute, for each $\ell \in \llbracket L \rrbracket$,

$$\text{MSE}_\ell^\bullet := \frac{1}{N} \sum_{n \in \llbracket N \rrbracket} (y'_{n,\ell} - \hat{y}'_{n,\ell}^\bullet)^2 \quad (4.10)$$

where we substitute $\text{SL}_1, \text{SL}_2, \text{OT}, \text{HYB}$ for the symbol \bullet . The average and standard deviations of these numbers are reported in Table 4.2. There is no stark differences in terms of standard deviations. In terms of average, the estimates yielded by the OT-procedure outperform those obtained by super learning. However, it is the hybrid procedure that emerges as the top performer. Figure 4.2 allows us to go beyond comparisons in average. More than two thirds of the points are situated to the left of the black vertical line, meaning that $\text{MSE}_\ell^{\text{OT}}$ is smaller than $\text{MSE}_\ell^{\text{SL}_2}$ for the corresponding ℓ s. Likewise, 29 out of 30 blue points are situated below the horizontal black line, meaning that $\text{MSE}_\ell^{\text{HYB}}$ is smaller than $\text{MSE}_\ell^{\text{SL}_2}$ for the corresponding ℓ s, while 24 out of 30 red points are situated below the horizontal black line, meaning that $\text{MSE}_\ell^{\text{HYB}}$ is smaller than $\text{MSE}_\ell^{\text{OT}}$ for the corresponding ℓ s. In particular, the average pattern unveiled by Table 4.2 remains consistent even before averaging: the hybrid procedure exhibits superior performance, surpassing the OT-procedure, which in turn outperforms the second super learning procedure.

4.6 Forecasting the requests of the government declaration of natural disaster for a drought event in France

4.6.1 Fine-tuning the OT-procedure

Defining a cost function. To begin with, we address the challenge of defining a cost function $c : (\mathcal{X} \times \mathbb{R}) \times (\mathcal{X} \times \mathbb{R}) \rightarrow \mathbb{R}_+$ (4.4). In view of the description of a generic

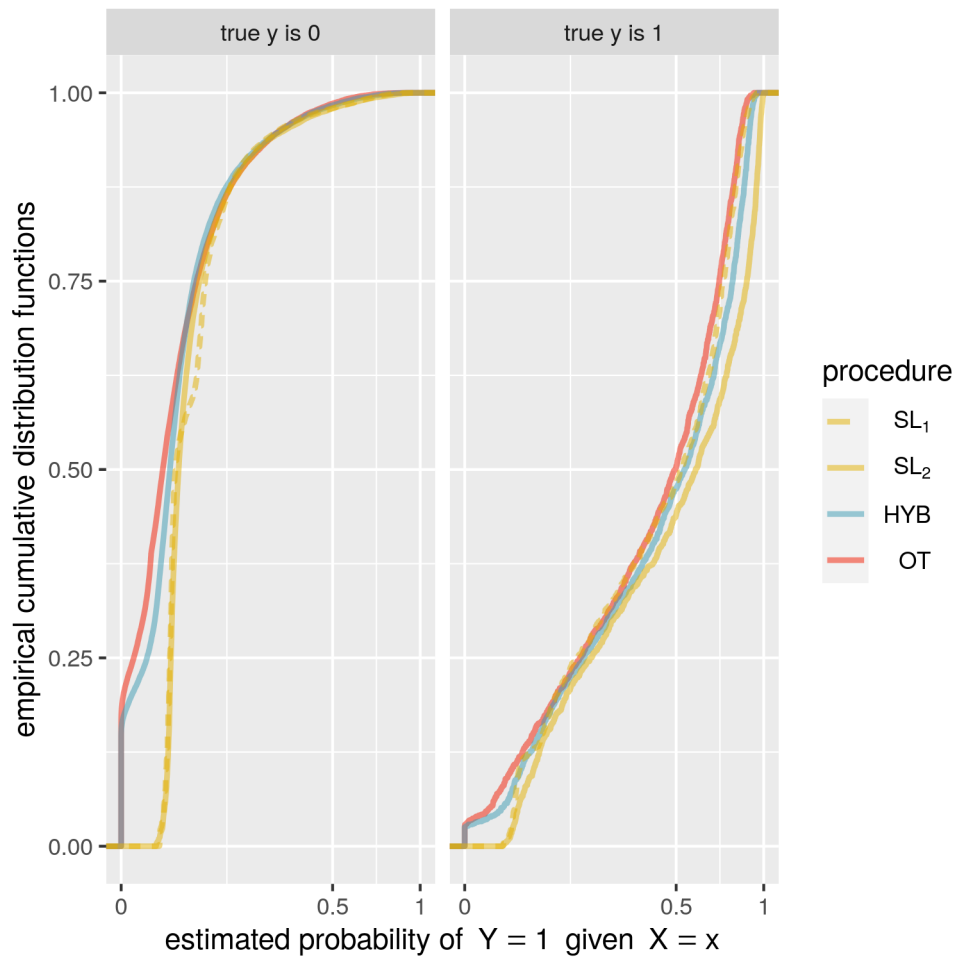


Figure 4.1: Empirical cumulative distribution functions of the sets $\{\hat{y}'_{n,\ell}^\bullet : \ell \in \llbracket L \rrbracket, n \in \llbracket N \rrbracket \text{ st } y'_{n,\ell} = y\}$ for $y = 0$ (left-hand side panel) and $y = 1$ (right-hand side panel), where the symbol \bullet stands for SL_1, SL_2, OT, HYB .

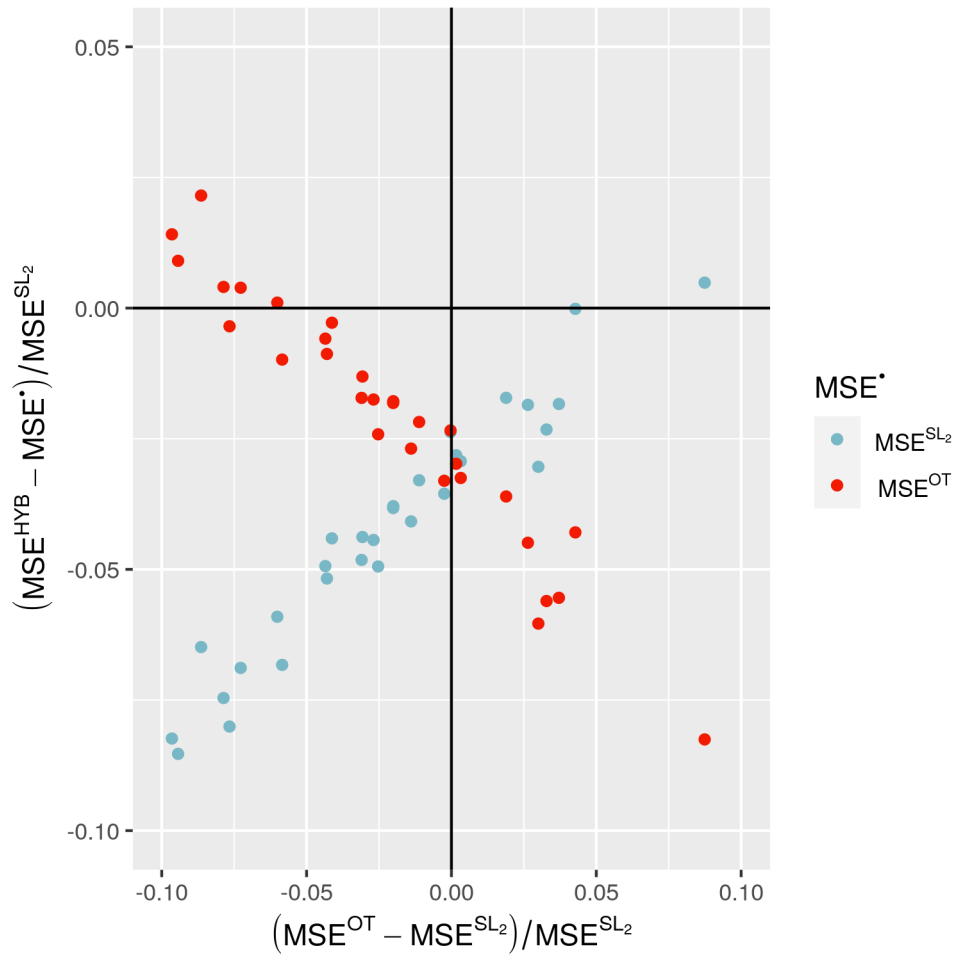


Figure 4.2: Scatterplot of $(MSE_\ell^{HYB} - MSE_\ell^*) / MSE_\ell^{SL_2}$ against $(MSE_\ell^{OT} - MSE_\ell^{SL_2}) / MSE_\ell^{SL_2}$ ($\ell \in \llbracket 30 \rrbracket$) where the symbol \bullet stands for SL_2 (blue) or OT (red). See also Table 4.2.

procedure	MSE	
	average	std. deviation
SL ₁	0.0361	0.0046
SL ₂	0.0345	0.0048
HYB	0.0330	0.0045
OT	0.0337	0.0044

Table 4.2: Averages and standard deviations of the mean squared errors $\{MSE_\ell^\bullet : \ell \in \llbracket L \rrbracket\}$ (4.10) where the symbol \bullet stands for SL_1, SL_2, OT, HYB and $L = 30$. See also Figure 4.2. In each column, the smallest value stands out in bold characters.

vector of covariates $x \in \mathcal{X}$ made in Section 4.2.1, let us rewrite $x := (x_{[1]}, \dots, x_{[4]})$ where $x_{[1]}, x_{[2]}, x_{[3]}$ and $x_{[4]}$ respectively regroup the covariates that collectively describe the corresponding city ($x_{[1]}$, 16 variables) and its exposure to drought events ($x_{[2]}$, 25 variables), provide a history of its past requests of declaration of natural disaster for a drought event, successful or not ($x_{[3]}$, 13 variables), and describe the city's vicinity ($x_{[4]}$, 13 variables).

Let $\bar{\xi}_1$ and std_1 be the vectors whose components are the component-specific mean and standard deviation of $\{\xi_{\alpha,1,u} : \alpha \in \mathcal{A}_1, u \in \mathcal{U}_1, \zeta_{\alpha,1,u} = 0\} \subset \mathcal{X}$, that is, the set of covariates corresponding to year 2019, and let $\bar{\zeta}_1$ be the $\|\cdot\|_1$ -norm of $\{\zeta_{\alpha,1} : \alpha \in \mathcal{A}_1\}$, that is, the number of cities which made a request for year 2019. For any generic vector of covariates $x \in \mathcal{X}$, denote (using the entrywise division of vectors)

$$\tilde{x} := \frac{x - \bar{\xi}_1}{\text{std}_1}. \quad (4.11)$$

We select a cost function in the parametric set $\{c_a : a \in \mathbb{R}_+^5\}$ where, for any $a \in \mathbb{R}_+^5$ and $x, x' \in \mathcal{X}, y, y' \in \mathbb{R}$,

$$c_a((x, y), (x', y')) := \sum_{k=1}^4 a_k \|\tilde{x}_{[k]} - \tilde{x}'_{[k]}\|_2^2 + a_5 (y - y')^2. \quad (4.12)$$

To do so, we rely on HYPERBAND, an algorithm which reformulates hyperparameter optimization as a pure-exploration, adaptive resource allocation problem addressing how to allocate resources among randomly generated hyperparameter configurations [Li et al., 2018]. Specifically, in view of 4.3, we set $\gamma = 10^{-2}$, $g_\tau : \theta \mapsto \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0, 1]^N\}$ with $\tau = \bar{\zeta}_1$ and, in view of (4.5) and Algorithm 1 in Section 4.4.3, we set

$$\begin{aligned} & \{(x_m, y_m) : m \in \llbracket M \rrbracket\} \\ & = \{(\xi_{\alpha,1,75}, \zeta_{\alpha,1}) : \alpha \in \mathcal{A}_1 \text{ st } \zeta_{\alpha,1,75} = 0\}, \end{aligned} \quad (4.13)$$

$$\begin{aligned} & \{x'_n : n \in \llbracket N \rrbracket\} \\ & = \{\xi_{\alpha,2,85} : \alpha \in \mathcal{A}_2 \text{ st } \zeta_{\alpha,2,85} = 0\}, \end{aligned} \quad (4.14)$$

$\alpha = 10^{-3}$, $\beta = 10^{-4}$ and $B = 128$. In words, setting (4.13) and (4.14) means that we exploit the data associated with the last week relative to year 2019 (that is, the (75 –

52) = 23rd week of 2020) to predict which cities will submit a request for the government declaration of natural disaster for a drought event for year 2020 during the last week relative to year 2020 (that is, the $(85 - 52) = 33$ rd week of 2021). As for the random generation of configurations $a = (a_1, a_2, a_3, a_4, a_5) \in \mathbb{R}_+^5$, we sample independently a_5 uniformly on $[1/5, 10]$ and (a_1, a_2, a_3, a_4) from the law of $73 \times \exp(Z) / \|\exp(Z)\|_1$ with Z drawn in \mathbb{R}^4 from the centered Gaussian law with identity covariance matrix and where the exponential is applied elementwise.

Moreover, in view of [Li et al., 2018, Algorithm 1, page 8], we set the maximum amount of resource that can be allocated to a single configuration (that is, the maximum number of iterations in Algorithm 1 that can be allocated to a randomly generated candidate $a \in \mathbb{R}_+^5$) to $R = 3000$ and the parameter controlling the proportion of configurations discarded in each round of SUCCESSIVEHALVING to $\eta = 10$. For this specific couple (R, η) , HYPERBAND consists of 4 independent “brackets” which we present in Table 4.3. In the bracket indexed by $s = 0$, $n_{0,0} = 4$ different $a \in \mathbb{R}_+^5$ (that is, configurations) are independently randomly generated; then each is allocated $r_{0,0} = 3000$ iterations in Algorithm 1 and associated with a score, a notion that we will clarify in the next paragraph. In the brackets indexed by $s \in \{1, 2, 3\}$, $n_{s,0}$ different $a \in \mathbb{R}_+^5$ are independently randomly generated; then, each is allocated $r_{s,0}$ iterations of Algorithm 1 and associated with a score. Next, recursively for $i = 1, \dots, s$, each of the $n_{s,i}$ configurations with the smallest scores is allocated $r_{s,i}$ iterations of Algorithm 1 and associated with a new score.

		brackets							
		$s = 3$		$s = 2$		$s = 1$		$s = 0$	
i		$n_{3,i}$	$r_{3,i}$	$n_{2,i}$	$r_{2,i}$	$n_{1,i}$	$r_{1,i}$	$n_{0,i}$	$r_{0,i}$
0		1000	3	134	30	20	300	4	3000
1		100	30	13	300	2	3000		
2		10	300	1	3000				
3		4	3000						

Table 4.3: Resource allocations and numbers of configurations $((r_{s,i}, n_{s,i}), i \in \{0, \dots, s\})$ in each bracket $s \in \{0, 1, 2, 3\}$ of the HYPERBAND procedure.

It only remains to clarify what are the aforementioned scores. For any configuration a randomly generated and tested while running HYPERBAND, let us denote by $\widehat{\zeta}_{\alpha,2}^{\text{OT},85}(a)$ the predicted probability output by Algorithm 1 that city α will eventually submit a request for the government declaration of natural disaster for a drought event for year 2020 for every $\alpha \in \mathcal{A}_2$ such that $\zeta_{\alpha,2,85} = 0$. The score associated with a is the MSE score

$$\frac{1}{N} \sum_{\alpha \in \mathcal{A}_2} (\widehat{\zeta}_{\alpha,2}^{\text{OT},85}(a) - \zeta_{\alpha,2})^2 \mathbf{1}\{\zeta_{\alpha,2,85} = 0\}. \quad (4.15)$$

This completes the description of the HYPERBAND algorithm that we run to select a cost function of the form (4.12). Eventually, we select the cost function c_a with $a \approx (16.75, 18.74, 30.57, 6.94, 0.34)$ (entries rounded to two decimal places).

Relative importance of the four groups of covariates concerning the selected cost function. To discuss the relative importance of each term in (4.12) with this choice of a , we sample uniformly without replacement $M = B = 128$ elements $x_1, \dots, x_m, \dots, x_M$ from $\{\xi_{\alpha,1,44} : \alpha \in \mathcal{A}\} \subset \mathcal{X}$ and, independently, $N = B = 128$ elements $x'_1, \dots, x'_n, \dots, x'_N$ from $\{\xi_{\alpha,2,48} : \alpha \in \mathcal{A}\}$ (recall that $\min \mathcal{U}_1 = 44$ and $\min \mathcal{U}_2 = 48$). In view of (4.11), each x_m yields $\tilde{x}_{m,[1]}, \tilde{x}_{m,[2]}, \tilde{x}_{m,[3]}, \tilde{x}_{m,[4]}$ and each x'_n yields $\tilde{x}'_{n,[1]}, \tilde{x}'_{n,[2]}, \tilde{x}'_{n,[3]}, \tilde{x}'_{n,[4]}$. We then compute the quartiles of the sets $\{\|\tilde{x}_{m,[k]} - \tilde{x}'_{n,[k]}\|_2^2 : m \in \llbracket M \rrbracket, n \in \llbracket N \rrbracket\}$ ($k = 1, 2, 3, 4$), which we report in Table 4.4.

Looking at Table 4.4 it seems that, for any $x, x' \in \mathcal{X}$ viewed as two cities' vectors of covariates, the sum $\sum_{k=1}^4 a_k \|x_{[k]} - x'_{[k]}\|_2^2$ (the left-hand side sum in (4.12)) is mainly driven, in decreasing order of importance, by $x_{[2]}, x'_{[2]}$ (the groups of 25 covariates describing the cities' exposures to drought events), $x_{[3]}, x'_{[3]}$ (the groups of 13 covariates describing the cities' histories of requests of declaration of natural disaster for a drought event), $x_{[1]}, x'_{[1]}$ (the groups of 16 covariates describing the cities) and $x_{[4]}, x'_{[4]}$ (the groups of 13 covariates describing the cities' vicinities). This is confirmed by Figure 4.3.

Figure 4.3 represents the cumulative distribution functions of the sets $\{\text{cst}_{m,n} \times a_k \|\tilde{x}_{m,[k]} - \tilde{x}'_{n,[k]}\|_2^2 : m, n \in \llbracket 128 \rrbracket\}$ ($k = 1, 2, 3, 4$) where each $\text{cst}_{m,n}$ (any $m, n \in \llbracket 128 \rrbracket$) is defined as

$$\text{cst}_{m,n} := \left(\sum_{k=1}^4 a_k \|\tilde{x}_{m,[k]} - \tilde{x}'_{n,[k]}\|_2^2 \right)^{-1}.$$

The more a cumulative distribution function is shifted to the right the more a generic sum $\sum_{k=1}^4 a_k \|x_{[k]} - x'_{[k]}\|_2^2$ (for any $x, x' \in \mathcal{X}$, the left-hand side sum in (4.12)) is driven by the corresponding groups of covariates. By this criterion, we recover the ordering suggested by Table 4.4.

covariates describing: ($\tilde{x}_{[k]}$)	a city ($k = 1$)	its exposure to drought events ($k = 2$)	its request history ($k = 3$)	its vicinity ($k = 4$)
minimum	0.40	2.80	2.01	0.00
1st quartile	5.25	7.41	2.01	1.26
median	6.20	8.75	3.69	2.35
3rd quartile	7.25	10.22	6.20	3.83
maximum	15.94	20.78	15.80	20.18
a	16.75	18.74	30.57	6.94

Table 4.4: Quartiles of the sets $\{\|\tilde{x}_{m,[k]} - \tilde{x}'_{n,[k]}\|_2^2 : m, n \in \llbracket 128 \rrbracket\}$ ($k = 1, 2, 3, 4$) where $\tilde{x}_1, \dots, \tilde{x}_{128}$ and $\tilde{x}'_1, \dots, \tilde{x}'_{128}$ are derived from x_1, \dots, x_{128} and x'_1, \dots, x'_{128} which are independently sampled, uniformly without replacement, from $\{\xi_{\alpha,1,44} : \alpha \in \mathcal{A}\}$ and $\{\xi_{\alpha,2,48} : \alpha \in \mathcal{A}\}$. The last row recalls the four first entries of a selected based on the HYPERBAND algorithm. See also Figure 4.3.

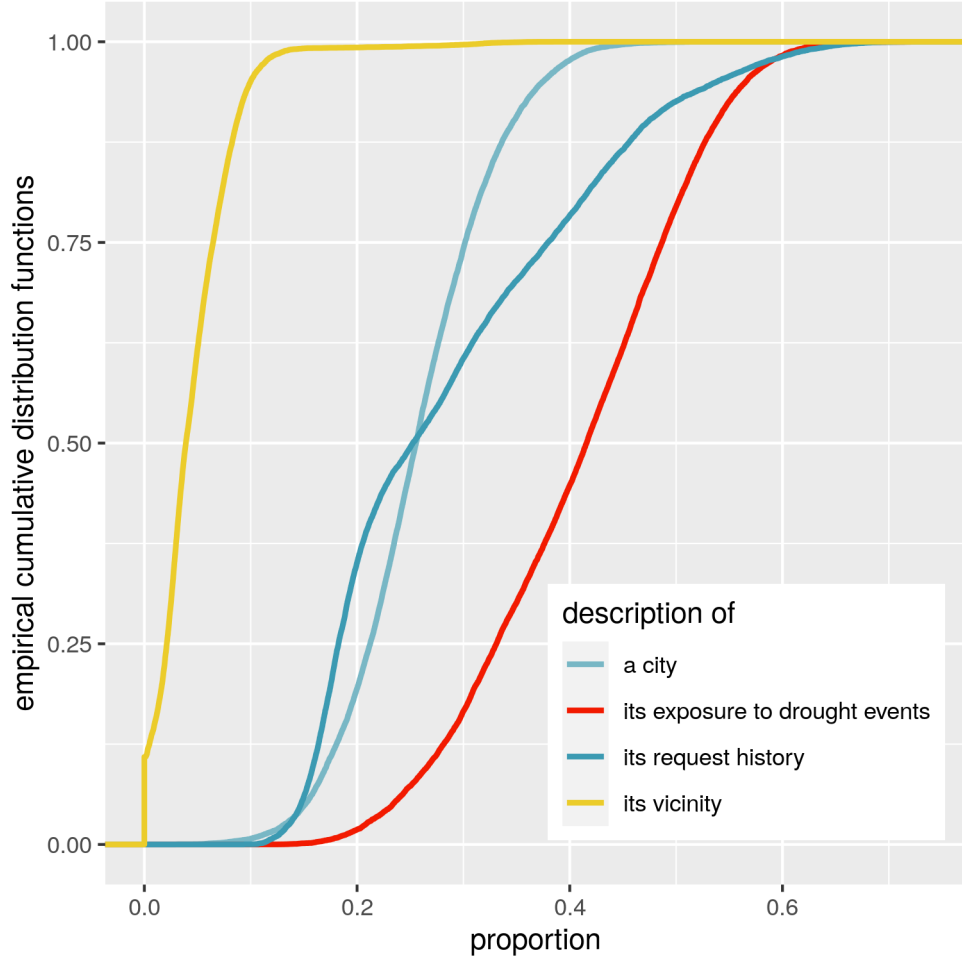


Figure 4.3: Cumulative distribution functions of the sets $\{cst_{m,n} \times a_k \|\tilde{x}_{m,[k]} - \tilde{x}'_{n,[k]}\|_2^2 : m, n \in \llbracket 128 \rrbracket\}$ ($k = 1, 2, 3, 4$) where $\tilde{x}_1, \dots, \tilde{x}_{128}$ and $\tilde{x}'_1, \dots, \tilde{x}'_{128}$ are derived from x_1, \dots, x_{128} and x'_1, \dots, x'_{128} which are independently sampled, uniformly without replacement, from $\{\xi_{\alpha,1,44} : \alpha \in \mathcal{A}\}$ and $\{\xi_{\alpha,2,48} : \alpha \in \mathcal{A}\}$, where a is selected based on the HYPERBAND algorithm, and where each $cst_{m,n}$ is such that $cst_{m,n} \times \sum_{k=1}^4 a_k \|\tilde{x}_{m,[k]} - \tilde{x}'_{n,[k]}\|_2^2 = 1$ for all $m, n \in \llbracket 128 \rrbracket$. The more a cumulative distribution function is shifted to the right the more a generic sum $\sum_{k=1}^4 a_k \|x_{[k]} - x'_{[k]}\|_2^2$ (for any $x, x' \in \mathcal{X}$, the left-hand side sum in (4.12)) is driven by the corresponding groups of covariates. See also Table 4.4.

Setting the remaining hyperparameters. Once the cost function is defined, we carry out a grid search to select values for γ (the regularization parameter in (4.3)), α and β (the learning rate and momentum parameters in Algorithm 1), with

$$(\gamma, \alpha, \beta) \in \{10^{-2}, 10^{-1}, 1\} \times \{10^{-3}, 5 \times 10^{-3}\} \times \{10^{-4}, 5 \times 10^{-4}\}.$$

For each possible triplet (γ, α, β) , we run Algorithm 1 with $g_\tau : \theta \mapsto \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0, 1]^N\}$ where $\tau = \bar{\zeta}_1$, (4.13), (4.14), $B = 128$ and collect the predicted probability $\hat{\zeta}_{\alpha,2}^{\text{OT},85}(\gamma, \alpha, \beta)$ that city α will eventually submit a request for the government declaration of natural disaster for a drought event for year 2020 for every $\alpha \in \mathcal{A}_2$ such that $\zeta_{\alpha,2,85} = 0$. The score associated with (γ, α, β) is the MSE score defined as in (4.15) with $\hat{\zeta}_{\alpha,2}^{\text{OT},85}(\gamma, \alpha, \beta)$ substituted for $\hat{\zeta}_{\alpha,2}^{\text{OT},85}(a)$. We select the triplet whose score is the smallest: $(\gamma, \alpha, \beta) = (10^{-2}, 10^{-3}, 10^{-4})$.

4.6.2 Alternative, classification-based approaches

As in the simulation study presented in Section 4.5, we also develop an alternative approach to predicting the requests of the government declaration of natural disaster for a drought event. We consider four individual algorithms in order to learn to classify each x'_n ($n \in \llbracket N \rrbracket$) using $\{(x_m, y_m) : m \in \llbracket M \rrbracket\}$. From a probabilistic viewpoint, the first algorithm, CST, approximates the conditional probability that $Y = 1$ given X under the form of a constant function (in X); the second algorithm, GLM, learns which element in a linear working model best approximates it (see `stats::glm`); the third algorithm, RANGER, approximates it under the form of a random forest (see `ranger::ranger`); the fourth algorithm, KNN, uses the nearest labelled neighbors of any x to estimate the conditional probability at $X = x$. More specifically, the linear working model at the core of GLM regresses Y linearly onto each component of X , treating as categorical variables the covariates characterizing a city's seismic and climatic zones, and uses a logit link function. RANGER uses the Gini splitting rule while the other hyperparameters are set to their default values specified in `ranger::ranger` [Wright and Ziegler, 2017]. As for KNN, it relies on the python class `sklearn.neighbors.KNeighborsClassifier` [Buitinck et al., 2013] and uses $k = 100$ neighbors, uniform weights, the ball tree algorithm [Liu et al., 2006, to handle the large learning data set] with a leaf size set to 30 and the weighted Euclidean $(x, x') \mapsto \|\tilde{x} - \tilde{x}'\|_2$.

We adopt a sequential learning viewpoint. Firstly, we train the four algorithms using all the data relative to year 2019, that is

$$\begin{aligned} & \{(x_m, y_m) : m \in \llbracket M \rrbracket\} \\ & = \{(\xi_{\alpha,1,u}, \zeta_{\alpha,1}) : \alpha \in \mathcal{A}_1, u \in \mathcal{U}_1 \text{ st } \zeta_{\alpha,1,u} = 0\}, \end{aligned}$$

yielding four functions $\hat{\zeta}_1^\bullet : \mathcal{X} \rightarrow [0, 1]$, where the symbol \bullet stands for CST, GLM, RANGER or KNN. Secondly, for each algorithm in turn, we compute the predicted probabilities of submitting a request relative to year 2020 for every week $u \in \mathcal{U}_2$ and all cities which did not submit a request yet by week u , that is $\hat{\zeta}_{\alpha,2}^{\bullet,u} := \hat{\zeta}_1^\bullet(\xi_{\alpha,2,u})$ for every $u \in \mathcal{U}_2$ and $\alpha \in \mathcal{A}_2$ such that $\zeta_{\alpha,2,u} = 0$. Thirdly, for each algorithm in turn, we compute the overall MSE score

$$\frac{\sum_{u \in \mathcal{U}_2} \sum_{\alpha \in \mathcal{A}_2} (\hat{\zeta}_{\alpha,2}^{\bullet,u} - \zeta_{\alpha,2})^2 \mathbf{1}\{\zeta_{\alpha,2,u} = 0\}}{\sum_{u \in \mathcal{U}_2} \sum_{\alpha \in \mathcal{A}_2} \mathbf{1}\{\zeta_{\alpha,2,u} = 0\}}.$$

The top-performing algorithm, GLM, is defined as the one with the smallest overall MSE score among all. We refer to it as the *discrete* super learner SL for year 2021. Lastly we retrain GLM, leveraging all data relative to years 2019 and 2020, that is

$$\begin{aligned} & \{(x_m, y_m) : m \in \llbracket M \rrbracket\} \\ & = \{(\xi_{\alpha,t,u}, \zeta_{\alpha,t}) : t = 1, 2, \alpha \in \mathcal{A}_t, u \in \mathcal{U}_t \text{ st } \zeta_{\alpha,t,u} = 0\}, \end{aligned}$$

yielding the function $\widehat{\zeta}_{1:2}^{\text{SL}} : \mathcal{X} \rightarrow [0, 1]$.

4.6.3 Results

We compute the predicted probabilities of submitting a request relative to year 2021 for every week $u \in \mathcal{U}_3$ and all cities which did not submit a request yet by week u , that is $\widehat{\zeta}_{\alpha,3}^{\text{SL},u} := \widehat{\zeta}_{1:2}^{\text{SL}}(\xi_{\alpha,3,u})$ for every $u \in \mathcal{U}_3$ and $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. Moreover, we run Algorithm 1 sequentially for each $u \in \mathcal{U}_3$, using the cost function (4.12) with $a \approx (16.75, 18.74, 30.57, 6.94, 0.34)$, $(\gamma, \alpha, \beta) = (10^{-2}, 10^{-3}, 10^{-4})$, $B = 128$, $T = 30,000$ and $g_\tau : \theta \mapsto \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0, 1]^N\}$ with $\tau = \|(\widehat{\zeta}_{\alpha,3}^{\text{SL},u})_{\alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u}=0}\|_1$. This yields the predictions $\widehat{\zeta}_{\alpha,3}^{\text{OT},u}$ for every $u \in \mathcal{U}_3$ and $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. Finally, we compute the predictions according to the hybrid procedure, that is, $\widehat{\zeta}_{\alpha,3}^{\text{HYB},u} := (\widehat{\zeta}_{\alpha,3}^{\text{SL},u} \times \widehat{\zeta}_{\alpha,3}^{\text{OT},u})^{1/2}$ for every $u \in \mathcal{U}_3$ and $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. Of note, it necessarily holds by design that

$$\|(\widehat{\zeta}_{\alpha,3}^{\text{HYB},u})_{\alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u}=0}\|_1 \leq \|(\widehat{\zeta}_{\alpha,3}^{\text{SL},u})_{\alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u}=0}\|_1 \quad (4.16)$$

for every $u \in \mathcal{U}_3$. Indeed, for any $\theta, \theta' \in \mathbb{R}_+^N$ such that $\|\theta\|_1 \geq \|\theta'\|_1$, the Cauchy-Schwarz inequality yields

$$\|([\theta_n \theta'_n]^{1/2})_{n \in \llbracket N \rrbracket}\|_1 \leq (\|\theta\|_1 \times \|\theta'\|_1)^{1/2} \leq \|\theta\|_1.$$

Figure 4.4 shows on maps of France the probabilities $\{\widehat{\zeta}_{\alpha,3}^{\text{HYB},u} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ predicted by the hybrid procedure of submitting a request relative to year 2021 for weeks $u = 49$ and $u = 78$. It is worth emphasizing that there are no predicted probabilities within the range of 50% to 90% during week 78.

Figure 4.5 represents the ecdfs of the predicted probabilities $\{\widehat{\zeta}_{\alpha,3}^{\bullet,u} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ of submitting a request for the government declaration of natural disaster for a drought event for year 2021 output by the super learner, the OT-procedure and the hybrid procedure for a selection of weeks u : the 49th week of 2021 (December 6th to 12th, $u = \min \mathcal{U}_3 = 49$), the 7th, 17th and 26th weeks of 2022 (February 15th to 21st, $u = 59$; April 26th to May 2nd, $u = 69$; June 28th to July 4th, $u = \max \mathcal{U}_3 = 78$). For each week, the right-hand side and left-hand side panels respectively focus on cities that will and that will not submit a request eventually. As expected, the curves in the left-hand side panels dominate their counterparts in the right-hand side panels, illustrating the fact that the predicted probabilities are smaller (in law) for cities that will not submit a request eventually than for cities that will. The curves mainly differ around the origin. The left-hand side panels clearly showcase the ability of the OT-procedure to rightly

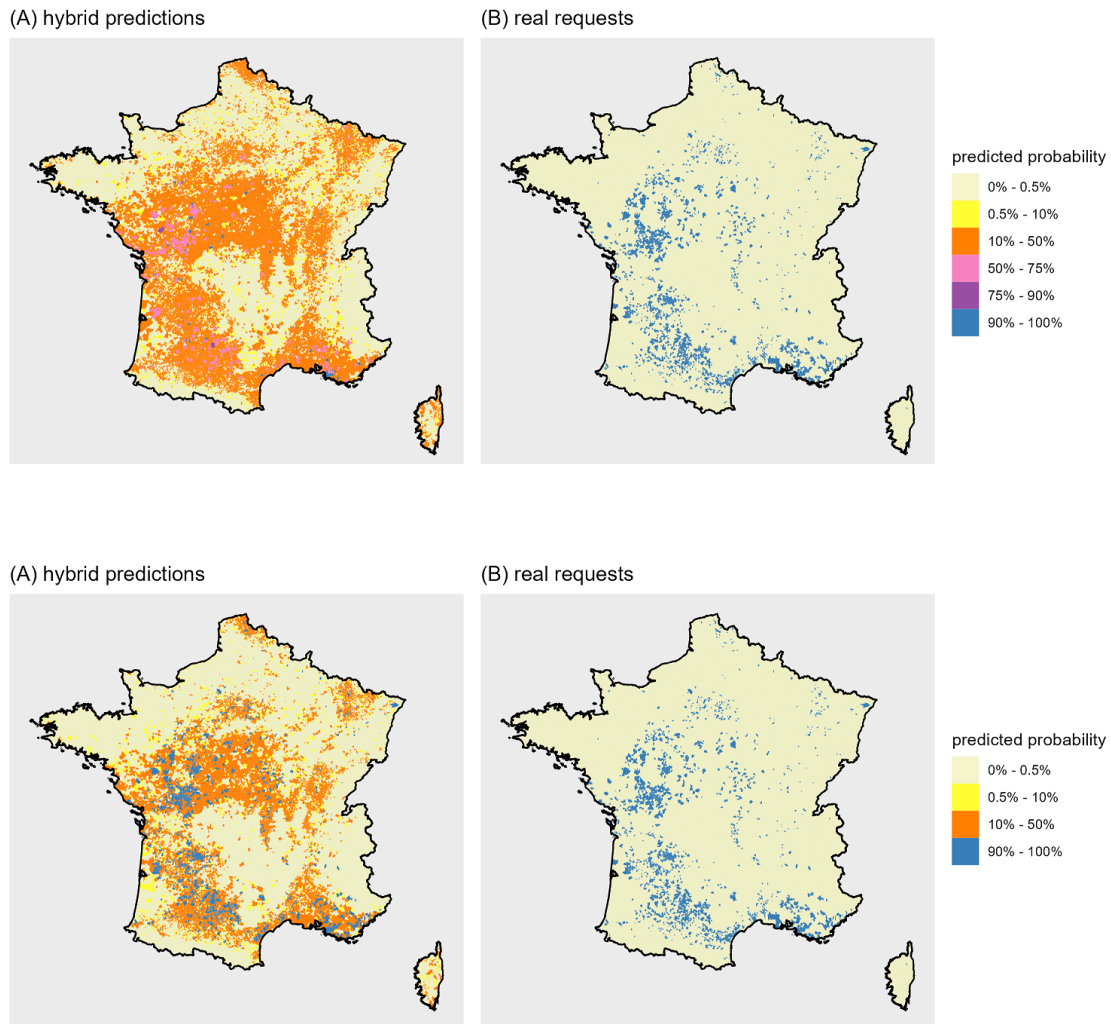


Figure 4.4: The left-hand side maps show the probabilities predicted by the hybrid procedure of submitting a request relative to year 2021 for weeks 49 (top) and 78 (bottom). The right-hand side maps show the cities that did submit a request eventually. In both left-hand side maps, the cities for which it was already known that they submitted a request are colored in blue. It is worth emphasizing that there are no predicted probabilities within the range of 50% to 90% during week 78.

assign a 0 probability to submit a request to cities that, indeed, will not submit one eventually: this concerns 49.5%, 51.2%, 50.7% and 56.4% of them for weeks 49, 59, 69 and 78 respectively. In contrast, the quantiles of order 49.5%, 51.2%, 50.7% and 56.4% of the super learner's predictions for these cities are 1.5%, 1.3%, 0.8% and 0.5% respectively. This notable ability comes at a price, as illustrated by the right-hand side panels showing that a 0-probability to submit a request is wrongly assigned to a fraction of the cities that, in fact, will submit one eventually: this concerns 4.3%, 7.6%, 6.7% and 14.6% of them for weeks 49, 59, 69 and 78 respectively. In comparison, the quantiles of order 4.3%, 7.6%, 6.7% and 14.6% of the super learner's predictions for these cities are 1.7%, 1.9%, 0.9% and 0.9% respectively.

Figure 4.6 compares the predicted probabilities of submitting a request for the government declaration of natural disaster for a drought event for year 2021 output by the super learner and by the OT-procedure during the 49th week of 2021 ($u = \min \mathcal{U}_3 = 49$) and the 26th week of 2022 ($u = \max \mathcal{U}_3 = 78$). For each week, the right-hand side and left-hand side panels respectively focus on cities that will and that will not submit a request eventually. Points lying above the first bisecting line correspond to cities $\alpha \in \mathcal{A}_3$ for which $\widehat{\zeta}_{\alpha,3}^{\text{OT},u} > \widehat{\zeta}_{\alpha,3}^{\text{SL},u}$. Colored points represent quantiles of order 10%, 50% and 90%. Two patterns emerge. On the one hand, for $u = 48$ and $u = 79$ both, when concentrating on cities that will not submit a request eventually: (a) the 10%-quantile and median of $\{\widehat{\zeta}_{\alpha,3}^{\text{OT},u} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ are smaller than those of $\{\widehat{\zeta}_{\alpha,3}^{\text{SL},u} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ while (b) the 90%-quantile of the former set is larger than that of the latter. Finding (a) is in favor of the OT-procedure while finding (b) is in favor of the super learner. On the other hand, for $u = 48$ and $u = 79$ both, when centering on cities that will submit a request eventually: (c) the median of $\{\widehat{\zeta}_{\alpha,3}^{\text{OT},u} : \alpha \in \mathcal{A}_3 \text{ st } (\zeta_{\alpha,3,u}, \zeta_{\alpha,3,u-}) = (1, 0)\}$ is larger than that of $\{\widehat{\zeta}_{\alpha,3}^{\text{SL},u} : \alpha \in \mathcal{A}_3 \text{ st } (\zeta_{\alpha,3,u}, \zeta_{\alpha,3,u-}) = (1, 0)\}$ while (d) the 10%- and 90%-quantiles of the former set are smaller than that of the latter. Finding (c) is in favor of the OT-procedure while finding (d) is in favor of the super learner.

Figure 4.7 pays special attention to the medians, representing those of the predicted probabilities of submitting a request for the government declaration of natural disaster for a drought event for year 2021 as output by the super learner, the OT-procedure and the hybrid procedure as weeks go by, its right-hand side and left-hand side panels focusing on cities that will and that will not submit a request eventually. A clear pattern emerges: when centering on cities that will not submit a request eventually, the week-specific median of the predictions made by the super learner is consistently larger than that of the predictions made by our procedure which, in turn, is consistently larger than that of the predictions made by the hybrid procedure. Conversely, when concentrating on cities that will submit a request eventually, the week-specific median of the predictions made by the super learner is consistently smaller than that of predictions made by the OT-procedure which, in turn, is consistently larger than that of the predictions made by the hybrid procedure. From this perspective, the hybrid procedure outperforms the OT-procedure which, in turn, performs better than the super learner.

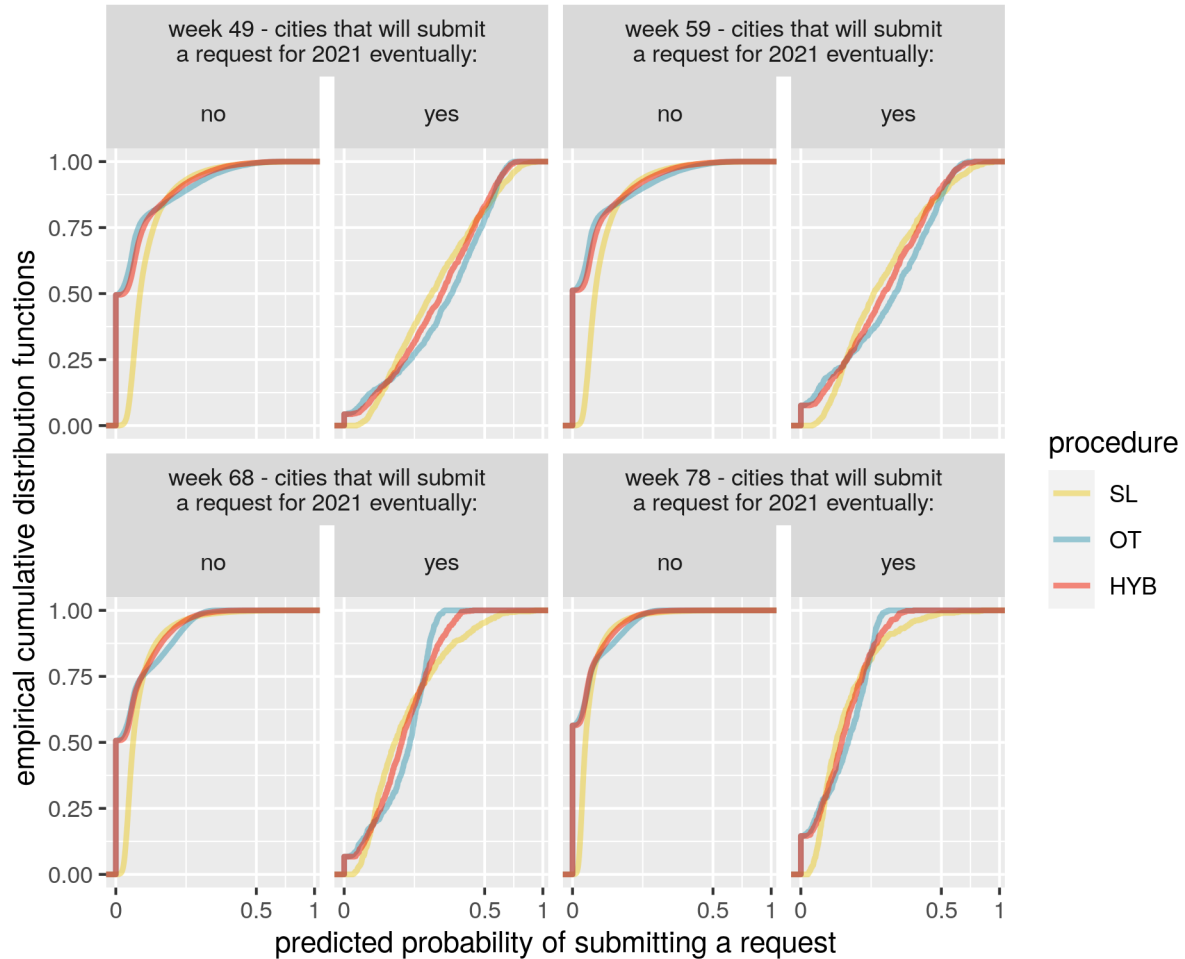


Figure 4.5: This plot shows, when week u is one of the 49th week of 2021 (December 6th to 12th), the $(59 - 52) = 7$ th, $(69 - 52) = 17$ th and $(78 - 52) = 26$ th weeks of 2022 (February 15th to 21st, April 26th to May 2nd, June 28th to July 4th), the empirical cumulative distribution functions (ecdfs) of the predicted probabilities of submitting a request made by procedures SL, OT and HYB separately for those cities that will not eventually submit a request for the government declaration of natural disaster for a drought event for year 2021 (that is, the ecdfs of $\{\hat{\zeta}_{\alpha,3}^{\bullet,u} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3} = 0\}$, left-hand side panels) and for those that will (that is, the ecdfs of $\{\hat{\zeta}_{\alpha,3}^{\bullet,u} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3} = 1\}$, right-hand side panels). See also Figure 4.7 for a focus on medians.

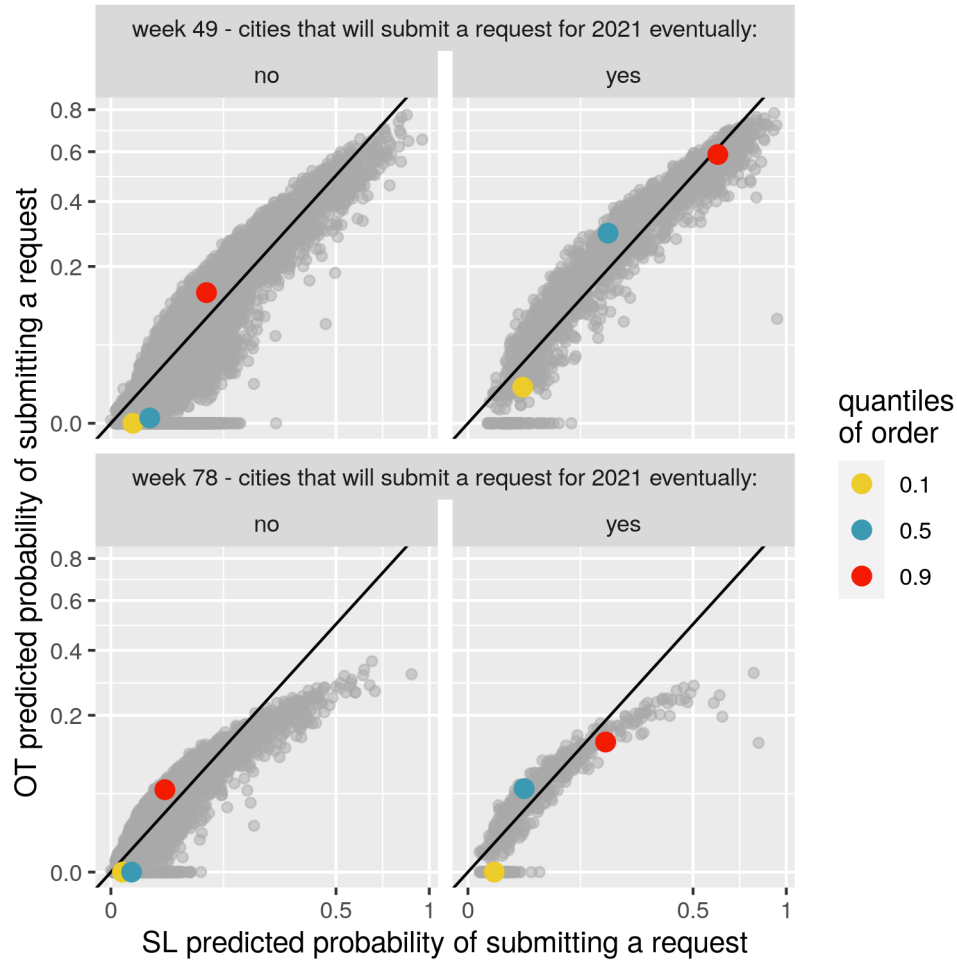


Figure 4.6: This plot shows, for week u equal either to the 49th week of 2021 (December 6th to December 12th) or the $(78 - 52) = 26$ th week of 2022 (June 27th to July 3rd), the predicted probabilities of submitting a request made by procedures SL (x -axis) and OT (y -axis) separately for those cities that will not eventually submit a request for the government declaration of natural disaster for a drought event for year 2021 (that is, $\{(\hat{\zeta}_{\alpha,3}^{SL,u}, \hat{\zeta}_{\alpha,3}^{OT,u}) : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 0\}$, left-hand side panels) and for those that will (that is, $\{(\hat{\zeta}_{\alpha,3}^{SL,u}, \hat{\zeta}_{\alpha,3}^{OT,u}) : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 1\}$, right-hand side panels). In addition, three colored points represent in each panel the coordinate-specific quantiles of order 10%, 50% and 90%.

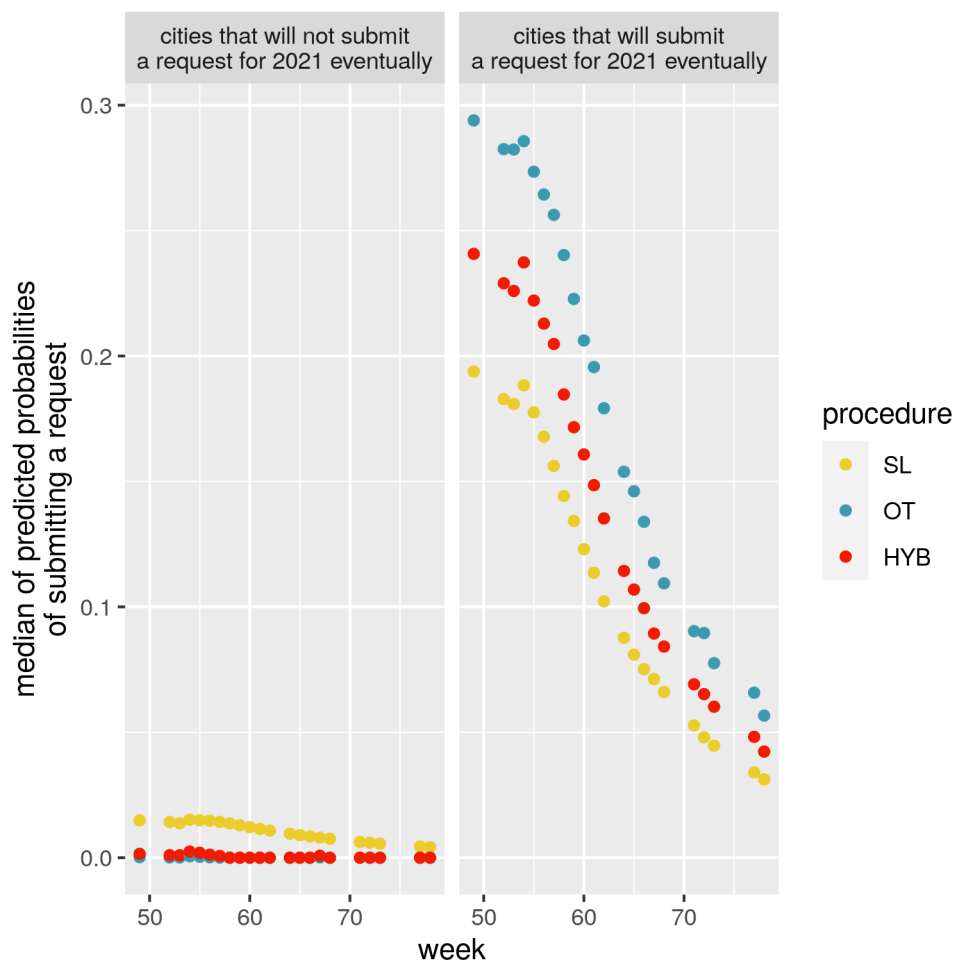


Figure 4.7: This plot shows, as week u goes from the 49th week of 2021 (December 6th to December 12th) to the $(78 - 52) = 26$ th week of 2022 (June 27th to July 3rd), the evolutions of the medians of the predicted probabilities of submitting a request made by procedures *SL*, *OT* and *HYB* separately for those cities that will not eventually submit a request for the government declaration of natural disaster for a drought event for year 2021 (that is, of $u \mapsto \text{median}\{\hat{\zeta}_{\alpha,3}^{\bullet,u} : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 0\}$, left-hand side panel) and for those that will (that is, of $u \mapsto \text{median}\{\hat{\zeta}_{\alpha,3}^{\bullet,u} : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 1\}$, right-hand side panel). See also Figure 4.5 for more comprehensive descriptions through empirical cumulative distribution functions.

To conclude, we report in Table 4.5 the week-specific MSE scores

$$\frac{\sum_{\alpha \in \mathcal{A}_3} (\hat{\zeta}_{\alpha,3}^{\bullet,u} - \zeta_{\alpha,3})^2 \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}}{\sum_{\alpha \in \mathcal{A}_3} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}} \quad (4.17)$$

(all $u \in \mathcal{U}_3$, the symbol \bullet standing for SL, OT and HYB). The key insight from Table 4.5 is that the hybrid procedure exhibits superior performance, by consistently outperforming both the OT-procedure and the super learner. Interestingly we also observe that, for every procedure, (4.17) decreases as $u \in \mathcal{U}_3$ increases, suggesting that the challenge of forecasting which cities will eventually request the government declaration of natural disaster for a drought event becomes progressively less challenging as the weeks go by. The evolution of (4.17) for $u \in \mathcal{U}_3$ is represented in Figure 4.8, with those of the stock of requests already submitted ($u \mapsto \sum_{\alpha \in \mathcal{A}_3} \zeta_{\alpha,3,u}$, necessarily increasing) and of the sum of the predicted probabilities that the cities which have not yet submitted such a request will eventually do, according to the hybrid procedure ($u \mapsto \sum_{\alpha \in \mathcal{U}_3} \hat{\zeta}_{\alpha,3}^{\text{HYB},u} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$). The quartiles and range of

$$\left\{ \sum_{\alpha \in \mathcal{A}_3} \zeta_{\alpha,3,u} + \sum_{\alpha \in \mathcal{U}_3} \hat{\zeta}_{\alpha,3}^{\text{HYB},u} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\} : u \in \mathcal{U}_3 \right\} \quad (4.18)$$

(the heights of the bars in Figure 4.8) are 1572 (minimum), 1636 (first quartile), 1731 (median), 1853 (third quartile), 1908 (maximum), 336 (range) while its mean is 1740. In comparison, the quartiles and range of

$$\left\{ \sum_{\alpha \in \mathcal{A}_3} \zeta_{\alpha,3,u} + \sum_{\alpha \in \mathcal{U}_3} \hat{\zeta}_{\alpha,3}^{\text{SL},u} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\} : u \in \mathcal{U}_3 \right\} \quad (4.19)$$

are 1662 (minimum), 1776 (first quartile), 1881 (median), 2051 (third quartile), 2133 (maximum), 471 (range), while its mean is 1905 – note that we could have substituted OT for SL in the above display. In view of (4.16), it was guaranteed that each of the quartile and mean associated to (4.18) would be smaller than its counterpart associated to (4.19). Both convex hulls of (4.18) and (4.19) contain the true value $\sum_{\alpha \in \mathcal{A}_3} \zeta_{\alpha,3} = 1696$, the former being more concentrated around it than the latter. This last observation stems from a comparison of the ranges of the sets and can be further substantiated by comparing the interquartile intervals, with that of (4.18) encompassing the true value, unlike that of (4.19).

4.6.4 On the importance of the variables used to make predictions

In this last subsection, we consider the influence that each covariate $\xi_{\alpha,3,u,s}$ (note the additional subscript s , indicating the s th covariate) in a generic $\xi_{\alpha,3,u}$ has on the prediction $\hat{\zeta}_{\alpha,3}^{\text{HYB},u}$ that city $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$ will eventually submit a request for the government declaration of natural disaster for a drought event relative to year 2021 based on data available at week $u \in \mathcal{U}_3$. The question pertains to the definition and estimation

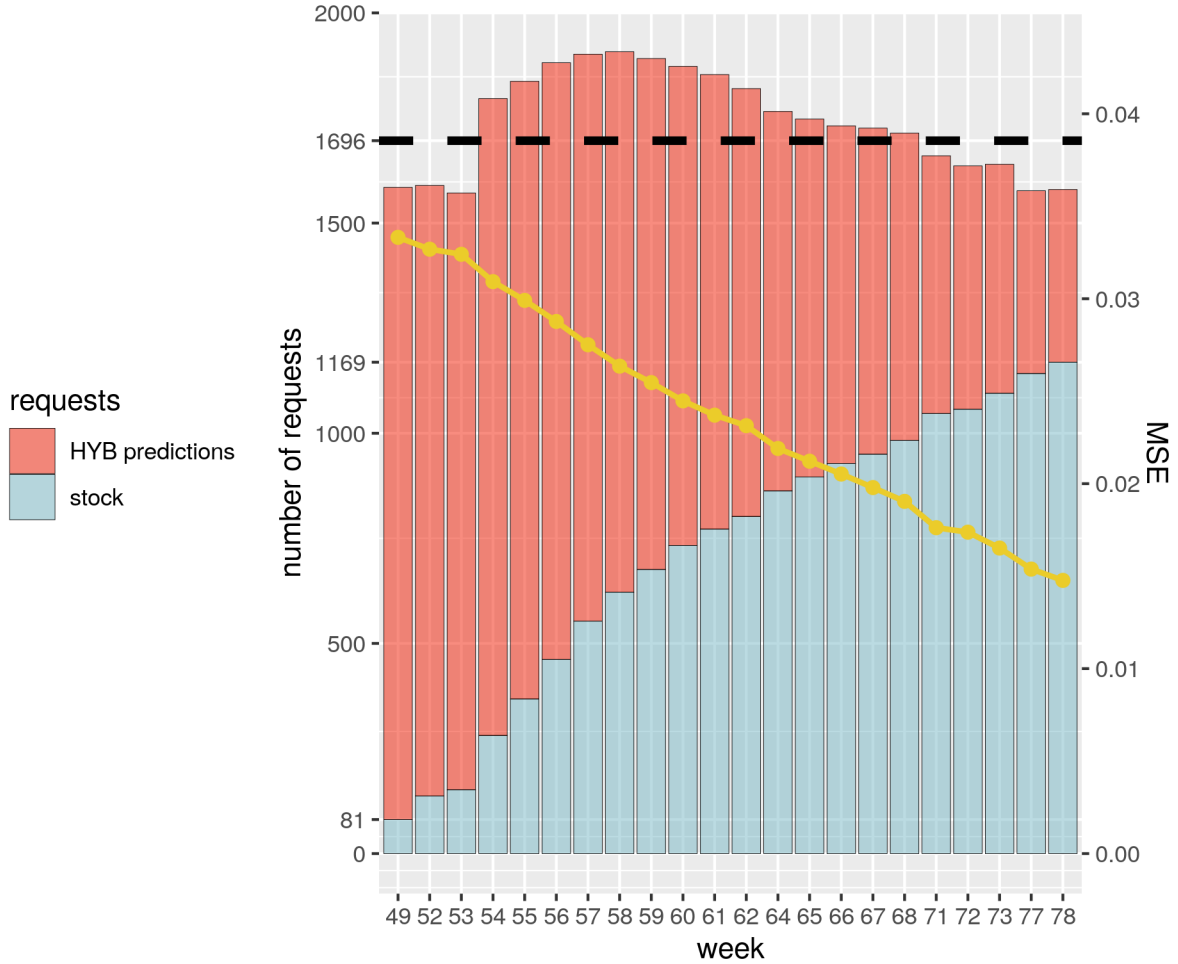


Figure 4.8: This plot shows, as week u goes from the 49th week of 2021 (December 6th to 12th) to the $(78 - 52) = 26$ th week of 2022 (June 27th to July 3rd), the evolutions of the cardinality of the stock of requests already submitted for the government declaration of natural disaster for a drought event for year 2021 (that is, of $u \mapsto \sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,3,u}$, in blue) and of the sum of the predicted probabilities that the cities which have not yet submitted such a request will eventually do (that is, of $u \mapsto \sum_{\alpha \in \mathcal{A}_t} \hat{\zeta}_{\alpha,3}^{HYB,u} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$, in red). The actual eventual number of such requests (that is, $\sum_{\alpha \in \mathcal{A}_t} \zeta_{\alpha,3}$, which equals 1696) is also represented (horizontal dashed line). In addition, the plot shows the evolution of MSE (that is, of $u \mapsto n_{3,u}^{-1} \sum_{\alpha \in \mathcal{A}_3} (\hat{\zeta}_{\alpha,3}^{HYB,u} - \zeta_{\alpha,3})^2 \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ where $n_{3,u} := \sum_{\alpha \in \mathcal{A}_3} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ is the number of cities which have not submitted such a request yet at week u , in yellow). See also Table 4.5.

week u	MSE			week u	MSE		
	SL	OT	HYB		SL	OT	HYB
49	0.0341	0.0341	0.0333	62	0.0236	0.0241	0.0231
52	0.0336	0.0333	0.0327	64	0.0223	0.0228	0.0219
53	0.0332	0.0331	0.0324	65	0.0216	0.0221	0.0212
54	0.0317	0.0321	0.0309	66	0.0208	0.0214	0.0205
55	0.0307	0.0311	0.0299	67	0.0202	0.0203	0.0198
56	0.0294	0.0302	0.0288	68	0.0195	0.0195	0.0190
57	0.0281	0.0290	0.0275	71	0.0179	0.0180	0.0176
58	0.0268	0.0280	0.0264	72	0.0177	0.0177	0.0174
59	0.0258	0.0271	0.0255	73	0.0168	0.0168	0.0165
60	0.0248	0.0261	0.0245	77	0.0156	0.0156	0.0154
61	0.0242	0.0248	0.0237	78	0.0150	0.0150	0.0148

Table 4.5: Evolution of MSE $u \mapsto n_{3,u}^{-1} \sum_{\alpha \in \mathcal{A}_3} (\hat{\zeta}_{\alpha,3}^{\bullet,u} - \zeta_{\alpha,3})^2 \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ where $n_{3,u} := \sum_{\alpha \in \mathcal{A}_3} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ is the number of cities which have not submitted such a request yet at week $u \in \mathcal{U}_3$ and the symbol \bullet stands for SL, OT, HYB. In each row, the smallest value stand out in bold characters. See also Figure 4.8.

of variable importance measures. The literature on this topic is rich, with notable contributions from studies such as [van der Laan, 2006, Hubbard et al., 2016, Williamson et al., 2021] on the one hand and [Lundberg and Lee, 2017, and references therein] on the other hand, offering valuable insights on how to tackle this question. However, applying these existing approaches to our specific scenario is impractical, mainly due to the interdependence of the data-structures specific to each $(\alpha, u) \in \mathcal{A}_3 \times \mathcal{U}_3$ and the fact that we are dealing with a relatively large number of covariates. As a result, we propose a simple approach tailored to the circumstances of the present situation. The approach is very similar to the one developed in [Ecoto and Chambaz, 2022a, Section 4.4].

Set arbitrarily $s \in \llbracket 67 \rrbracket$ and $u \in \mathcal{U}_3$.

- If s is such that the covariate $\xi_{\alpha,3,u,s}$ corresponds to the overall number of French cities that submitted a request for year 2021 during week u or before, or to the ratio of the logarithm of that overall number to u (two elements of the description of a city's request history), then we cannot quantify the covariate's importance because all cities $\alpha \in \mathcal{U}_3$ share a common value.
- If s is such that $\xi_{\alpha,3,u,s}$ ($\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$) take v values with $2 \leq v \leq 5$, then we let ρ_s be the correlation ratio computed based on $\{(\hat{\zeta}_{\alpha,3}^{\text{HYB},u}, \xi_{\alpha,3,u,s}) : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$:

$$\rho_s^u := \left(\frac{\sum_{\nu=1}^v n_\nu (\bar{\zeta}_\nu - \bar{\zeta})^2}{\sum_{\alpha \in \mathcal{A}_3} (\hat{\zeta}_{\alpha,3}^{\text{HYB},u} - \bar{\zeta})^2 \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}} \right)^{1/2}$$

where $\bar{\zeta}_\nu$ is the average of the $\hat{\zeta}_{\alpha,3}^{\text{HYB},u}$ s such that $\xi_{\alpha,3,u,s} = \nu$ and $\bar{\zeta}$ is the average of all $\hat{\zeta}_{\alpha,3}^{\text{HYB},u}$ s.

- Otherwise, we treat the covariate $\xi_{\alpha,3,u,s}$ ($\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$) as a continuous variable and let ρ_s^u be the absolute value of the Spearman rank correlation coefficient [Hollander and Wolfe, 1999, Section 8.5] computed based on $\{(\widehat{\zeta}_{\alpha,3}^{\text{HYB},u}, \xi_{\alpha,3,u,s}) : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$.

Note that, in the second case, we could have defined ρ_s^u as Wilcoxon test's statistic (case $v = 2$) or the Kruskal-Wallis test's statistics (case $3 \leq v \leq 5$) [see Hollander and Wolfe, 1999, Sections 3.1 and 6.1]. By guaranteeing that all ρ_s^u s naturally lie in $[0, 1]$, the present choice eases comparisons.

In all cases, the magnitude of ρ_s^u directly reflects the strength of the association between the sth covariate and the predictions made at week $u \in \mathcal{U}_3$. We resort to permutation tests to assess significance levels, with one million independent permutations drawn uniformly in each of the above cases. The maximum value obtained by permutation equals 3.16%.

Figure 4.9 shows the evolutions of $u \mapsto \rho_s^u$ for every eligible $s \in \llbracket 67 \rrbracket$, where the covariates are grouped based on the type of information they contribute. In each panel, values above the black horizontal lines (y -intercept at $(0, 3.16\%)$) are considered highly significant according to the permutation tests. From this perspective, most covariates play an effective role in the predictions. For the covariates related to a city's description, its exposure to drought events, or its request history, the curves appear relatively flat, indicating a steady strength of association with the predictions over time. In contrast, for the covariates describing a city's vicinity, the curves lying above the horizontal line show an increasing trend before levelling off. This suggests that the strength of association for each corresponding covariate gradually increases then stabilizes over time. In Table 4.6, we report the five variables which, in each group of covariates, feature the largest average variable importance ($\sum_{u \in \mathcal{U}_3} \rho_s^u / \text{card } \mathcal{U}_3$).

4.7 Discussion

This study is motivated by the challenging task of forecasting which cities in France will submit a request for the government declaration of natural disaster for a drought event. While the problem can be addressed as a classification task using standard classification algorithms, we take a slightly different perspective and introduce an alternative procedure based on optimal transport theory [Peyré and Cuturi, 2020] and iPiano [Ochs et al., 2015], an inertial proximal algorithm for nonconvex optimization.

We build the OT-procedure upon two core ideas. Firstly, we aim to predict whether a city will submit a request by making an interpretable comparison of the city's covariates with those of other cities whose submission status may be already known. Secondly, recognizing that relatively few cities will submit requests, we seek to control the sparsity of our predictions and encourage 0-predictions, indicating cases where we predict that a city will not submit a request. Additionally, we develop a hybrid procedure that synergistically combines and utilizes both types of predictions, derived from classification algorithms and the OT-procedure.

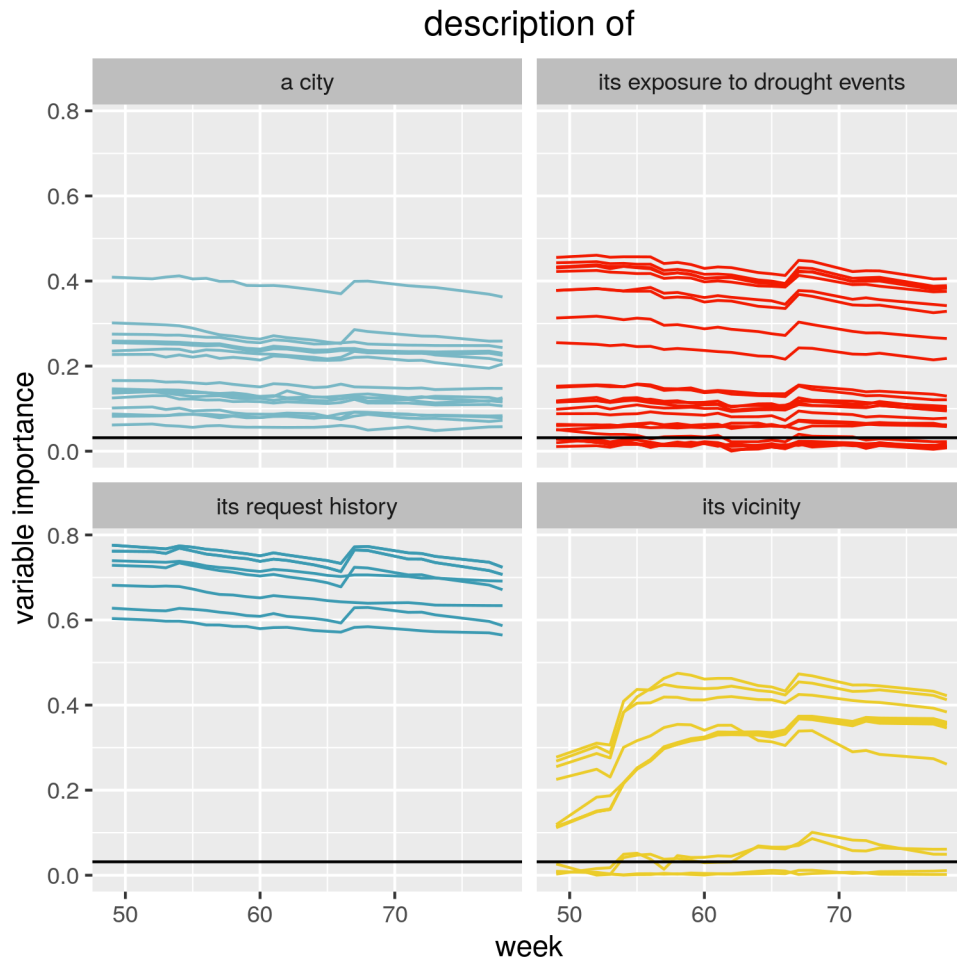


Figure 4.9: This plot shows, as week u goes from the 49th week of 2021 (December 6th to 12th) to the $(78 - 52) = 26$ th week of 2022 (June 27th to July 3rd), the evolutions of the importance of each variable used to make predictions, as defined in Section 4.6.4. For every eligible $s \in \llbracket 67 \rrbracket$, the larger is ρ_s^u , the stronger is the association between the s th covariate $\xi_{\alpha,3,u,s}$ and the prediction $\widehat{\zeta}_{\alpha,3}^{HYB,u}$ across $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. Values above the black horizontal lines are deemed highly significant based on permutation tests. See also Table 4.6.

description of	variable	avg. importance
a city	proportion of houses* in the 2nd clay-shrinkage-swelling hazard category	0.392
	climatic zone	0.275
	insured sum	0.259
	number of houses*	0.244
	population	0.239
its exposure to drought events	average SWI over Q1, Q2, Q3 [†]	0.436
	overall average SWI	0.420
	average SWI over Q2, Q3	0.412
	minimum SWI over Q2	0.412
	global minimum SWI	0.402
its request history	number of requests submitted during the 5 previous years	0.757
	number of requests submitted since 1990	0.744
	number of requests denied during the 2 previous years	0.715
	number of requests granted during the 2 previous years	0.708
	indicator of request denied the previous year	0.654
its vicinity	number of claims in the same department	0.423
	proportion of cities in the same department that submitted a request for year 2023 before week u	0.416
	proportion of cities in the same department that submitted a request for the first time during the 5 previous years	0.392
	ratio of the number of claims in the same department to the number of cities in the department	0.308
	number of neighboring cities that submitted a request for year 2023 before week u	0.305

* within the city's limits

[†] Q1, Q2, Q3, Q4 are the 1st to 4th quarters

Table 4.6: *The five variables used to make predictions with the highest average importance ($\sum_{u \in \mathcal{U}_3} \rho_s^u / \text{card} \mathcal{U}_3$, see definition in Section 4.6.4) in each group of covariates. For every eligible $s \in \llbracket 67 \rrbracket$, the larger is ρ_s^u , the stronger is the association between the s th covariate $\xi_{\alpha,3,u,s}$ and the prediction $\hat{\zeta}_{\alpha,3}^{HYB,u}$ across $\alpha \in \mathcal{A}_3$ such that $\zeta_{\alpha,3,u} = 0$. See also Figure 4.9.*

We develop and program an algorithm that hinges on iPiano and a mini-batch procedure to cope with large data sets, see Algorithm 1. The convergence of the iPiano algorithm is established, using the notion of \mathfrak{o} -minimal structures from the field of tame geometry [Wilkie, 1996] to prove that a critical function related to (4.3) satisfies the Kurdyka-Lojasiewicz property [Attouch et al., 2010]. Coded in `python/pytorch`, relying on the `GeomLoss` package [Feydy et al., 2019b] for its fast implementation of the Sinkhorn algorithm, the program is available at ♣.

We conduct a simulation study to illustrate the use of the OT-procedure and of the hybrid procedure in a simple context, laying the groundwork for the real-world application. The latter poses greater challenges than the former. Tangibly, these challenges arise because $\mathcal{X} \subset \mathbb{R}^d$ is a relatively high-dimensional space ($d = 67$) and because the sample sizes are large. Intangibly, the intricacies lie in the mechanisms that determine whether a request is submitted or not.

We rely on the HYPERBAND algorithm [Li et al., 2018] and on a simple grid search to define a relevant cost function and fine-tune the hyperparameters of Algorithm 1. An analysis of the cost function reveals that the more relevant groups of covariates are, in decreasing order of importance, the covariates related to a city’s exposure to drought events, its request history, its description and its vicinity.

For a total of 22 weeks spanning from the 49th week of 2021 (December 6th to 12th) to the 26th week of 2022 (June 28th to July 4th), intermittently, we predict whether or not the cities that have not yet submitted a request for the year 2021 will eventually do so. We employ the best of four standard classification algorithms, the OT-procedure and the hybrid procedure to make these predictions. Overall, the hybrid procedure yields enhanced forecasting accuracy, in particular while focusing on the estimation of the eventual number of requests.

For confidentiality reasons, we cannot compare our predictions to the predictions obtained by using the algorithm currently deployed at CCR. However, we were given the authorization to report the following fact. The average across the weeks of the MSE shown in column HYB of Table 4.5 is *more than 20%* smaller than the MSE of the predictions made by the algorithm currently deployed at CCR.

A simple analysis of the covariate’s importance sheds light on the strength of association between each covariate and the predictions. It suggests that most covariates play an effective role in the predictions.

We conclude by listing potential avenues for future research. Firstly, the procedures discussed in the study may benefit from the use of an enhanced version of the city-level SWI. By considering the variation in the nature of the soil across different regions of France, this refined version could contribute to making more accurate predictions. Secondly, to make the hybrid procedure more acceptable to the experts at CCR, it would be interesting to complement the analysis of the covariates’ importance. This additional analysis could offer further insights and explanations regarding the predictions. Thirdly, the current predictions obtained from the investigated procedures lack a measure of confidence. Developing a methodology to address this issue would be highly valuable. In conclusion, we acknowledge that the last two questions raised are very challenging,

notably due to the complex interdependence within the data set.

4.8 Appendix: checking the iPiano assumptions

The iPiano assumptions consist in

1. f being C^1 -smooth with a Lipschitz continuous gradient on $\text{dom } g_\tau$, see Section 4.8.1;
2. for any $\delta > 0$, $H_\delta : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ given by $H_\delta(\theta, \theta') := f(\theta') + g_\tau(\theta') + \delta \|\theta - \theta'\|_2^2$ having the Kurdyka-Lojasiewicz property at a cluster point (θ^*, θ^*) of the sequence $(\theta^k)_{k \geq 1}$, see Section 4.8.2.

4.8.1 The function f is C^1 -smooth and its gradient is Lipschitz continuous on $\text{dom } g_\tau$

4.8.1.1 Preliminaries

4.8.1.1.1 On matrix norms. For self-containedness, let us recall several definitions and results concerning matrix norms. For any matrix $A \in \mathbb{R}^{d \times d'}$, the Frobenius and maximum norms of A are given by $\|A\|_F := \left(\sum_{i \in \llbracket d \rrbracket, j \in \llbracket d' \rrbracket} A_{i,j}^2 \right)^{1/2}$ and $\|A\|_{\max} := \max\{|A_{i,j}| : i \in \llbracket d \rrbracket, j \in \llbracket d' \rrbracket\}$. For any vector $x \in \mathbb{R}^d$, the variation seminorm of x is defined as $\|x\|_{\text{var}} := \max\{x_i : i \in \llbracket d \rrbracket\} - \min\{x_i : i \in \llbracket d \rrbracket\}$. We will use the following classical inequalities and equality:

$$\forall A \in \mathbb{R}^{d \times d'}, \forall B \in \mathbb{R}^{d' \times d''}, \|AB\|_F \leq \|A\|_F \|B\|_F; \quad (4.20)$$

$$\forall A \in \mathbb{R}^{d \times d'}, \forall x \in \mathbb{R}^{d'}, \|Ax\|_2 \leq \|A\|_F \|x\|_2; \quad (4.21)$$

$$\forall x \in \mathbb{R}^d, \|\text{diag}(x)\|_F = \|x\|_2; \quad (4.22)$$

$$\forall x \in \mathbb{R}^d, \|x\|_{\text{var}} \leq 2\|x\|_\infty; \quad (4.23)$$

$$\forall x \in \{0\} \times \mathbb{R}^{d-1}, \|x\|_\infty \leq \|x\|_{\text{var}}. \quad (4.24)$$

4.8.1.1.2 On the Hilbert projective metric. The Hilbert projective metric on $(\mathbb{R}_+^*)^d$ is defined by

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') := \log \max \left\{ \frac{x_i x'_j}{x'_i x_j} : i, j \in \llbracket d \rrbracket \right\}.$$

We will use the following properties [Birkhoff, 1957]:

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') = \|\log(x) - \log(x')\|_{\text{var}}; \quad (4.25)$$

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') = d_{\mathcal{H}}(x/x', \mathbf{1}_d) = d_{\mathcal{H}}(\mathbf{1}_d/x', \mathbf{1}_d/x); \quad (4.26)$$

$$\forall K \in (\mathbb{R}_+^*)^{d \times d'}, \forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(Kx, Kx') \leq \lambda(K) d_{\mathcal{H}}(x, x'), \quad (4.27)$$

where $\lambda(K) := \frac{\sqrt{\eta(K)} - 1}{\sqrt{\eta(K)} + 1} < 1$ with $\eta(K) := \max \left\{ \frac{K_{i,k} K_{j,\ell}}{K_{j,k} K_{i,\ell}} : i, j \in \llbracket d \rrbracket, k, \ell \in \llbracket d' \rrbracket \right\}$.

We end this section with a lemma.

Lemma 8. *Let $x, x' \in (\mathbb{R}_+^*)^d$ be such that $0 < t \leq \min\{x_j, x'_j : j \in \llbracket d \rrbracket\} \leq \max\{x_j, x'_j : j \in \llbracket d \rrbracket\} \leq T$. It holds that $\frac{1}{2}td_{\mathcal{H}}(x, x') \leq \|x - x'\|_2$. Moreover, if $x_1 = x'_1 = 1$, then it also holds that $\|x - x'\|_2 \leq \sqrt{d}Td_{\mathcal{H}}(x, x')$.*

Proof. Set $x, x' \in (\mathbb{R}_+^*)^d$ as in the statement of the lemma, and denote $\ell := \log(x)$, $\ell' := \log(x')$ (the logarithms are elementwise). Set arbitrarily $i \in \llbracket d \rrbracket$. We can assume without loss of generality that $x_i \geq x'_i$ (or, equivalently, $\ell_i \geq \ell'_i$). Therefore if $x_1 = x'_1 = 1$ (or, equivalently, $\ell_1 = \ell'_1 = 0$), then

$$\begin{aligned} |x_i - x'_i| &= \max(x_i, x'_i) \times |1 - e^{-|\ell_i - \ell'_i|}| \\ &\leq T \times |\ell_i - \ell'_i| \quad \text{because } |1 - e^{-|q|}| \leq |q| \text{ for all } q \in \mathbb{R} \\ &\leq T \times \|\ell - \ell'\|_{\infty} \\ &\leq T \times \|\ell - \ell'\|_{\text{var}} \quad \text{by (4.24) since } \ell_1 = \ell'_1 = 0 \\ &= Td_{\mathcal{H}}(x, x') \quad \text{by (4.25)}. \end{aligned}$$

Consequently, $\|x - x'\|_2 \leq \sqrt{d}\|x - x'\|_{\infty} \leq \sqrt{d}Td_{\mathcal{H}}(x, x')$. Furthermore,

$$\begin{aligned} |x_i - x'_i| &= \min(x_i, x'_i) \times |e^{|\ell_i - \ell'_i|} - 1| \\ &\geq t \times |\ell_i - \ell'_i| \quad \text{because } |e^{|q|} - 1| \geq |q| \text{ for all } q \in \mathbb{R}. \end{aligned}$$

It follows that

$$\begin{aligned} \|x - x'\|_2 &\geq \|x - x'\|_{\infty} \geq t\|\ell - \ell'\|_{\infty} \geq \frac{1}{2}t\|\ell - \ell'\|_{\text{var}} \quad \text{by (4.23)} \\ &= \frac{1}{2}td_{\mathcal{H}}(x, x') \quad \text{by (4.25)}. \end{aligned}$$

This completes the proof. □

4.8.1.2 The function f is differentiable

To prove that f is differentiable, we rely on the following classical result [Danskin, 1966]:

Theorem 9 (Danskin's theorem, Proposition B.25 in Bertsekas [1999]). *Let $\mathcal{C} \subset \mathbb{R}^d$ be a compact set and $\phi : \mathbb{R}^d \times \mathcal{C} \rightarrow \mathbb{R}$ be a continuous function such that $\phi(\cdot, y)$ is convex for every $y \in \mathcal{C}$. The function $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ given by $\psi(x) := \max_{y \in \mathcal{C}} \phi(x, y)$ is convex. Moreover, if there exists a unique \hat{y} maximizing $\phi(x, \cdot)$ and if $\phi(\cdot, \hat{y})$ is differentiable, then ψ is differentiable at x and $\nabla\psi(x) = \nabla\phi(\cdot, \hat{y})|_x$.*

Let $\mathcal{C} = \Pi_{R, R'}$ (a compact set) and $\phi : \mathbb{R}^{R \times R'} \times \Pi_{R, R'} \rightarrow \mathbb{R}$ be given by $\phi(C, P) := -[\langle P, C \rangle - \gamma E(P)]$. The function ϕ is continuous and $\phi(\cdot, P)$ is convex for every $P \in \Pi_{R, R'}$. Therefore, by the above theorem, the function $\psi : \mathbb{R}^{R \times R'} \rightarrow \mathbb{R}$ given by $\psi(C) := \max_{P \in \Pi_{R, R'}} \phi(C, P) = -\mathcal{W}_{\gamma}(C)$ is convex. Moreover, for every $C \in \mathbb{R}^{R \times R'}$, there exists a unique \hat{P}_C such that $\psi(C) = \phi(C, \hat{P}_C)$ [Cuturi and Doucet, 2014, Proposition 4.3] and $\phi(\cdot, \hat{P}_C)$ is affine hence differentiable. Therefore, $C \mapsto \mathcal{W}_{\gamma}(C)$ is differentiable at every $C \in \mathbb{R}^{R \times R'}$ with a gradient given by $\nabla\mathcal{W}_{\gamma}(C) = \hat{P}_C$.

We use now that $f = f_a - \frac{1}{2}f_b + \text{constant}$ with $f_a, f_b : \mathbb{R}^N \rightarrow \mathbb{R}$ given by

$$f_a(\theta) := \mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z}'(\theta))) \quad \text{and} \quad f_b(\theta) := \mathcal{W}_\gamma(C(\mathbf{z}'(\theta), \mathbf{z}'(\theta)))$$

where the cost matrices $C(\mathbf{z}, \mathbf{z}'(\theta))$ and $C(\mathbf{z}'(\theta), \mathbf{z}'(\theta))$ are such that $(C(\mathbf{z}, \mathbf{z}'(\theta)))_{m,n} := \text{dis}(x_m, x'_n)^2 + (y_m - \theta_n)^2$ and $(C(\mathbf{z}'(\theta), \mathbf{z}'(\theta)))_{n,n'} := \text{dis}(x'_n, x'_{n'})^2 + (\theta_n - \theta_{n'})^2$. In view of the previous paragraph, and by the chain rule, f_a and f_b are thus differentiable at every $\theta \in \mathbb{R}^N$ with gradients

$$\nabla f_a(\theta) = 2\left(\frac{1}{N}\theta - \widehat{P}_\theta^\top y\right) \quad \text{and} \quad \nabla f_b(\theta) = 2\left(\frac{2}{N}\theta - (\widehat{Q}_\theta + \widehat{Q}_\theta^\top)\theta\right)$$

(\widehat{P}_θ and \widehat{Q}_θ are defined in (4.8) and (4.9)). Therefore f is differentiable at every $\theta \in \mathbb{R}^N$ and (4.7) follows straightforwardly.

4.8.1.3 \widehat{P}_θ and \widehat{Q}_θ are Lipschitz continuous (as functions of θ)

The fact that $\theta \mapsto \widehat{P}_\theta$ and $\theta \mapsto \widehat{Q}_\theta$ are Lipschitz continuous on $\text{dom } g_\tau$ is a consequence of the following lemma.

Lemma 10. *Let $\theta \mapsto C(\theta)$ be a bounded and Lipschitz continuous function from $[0, 1]^{R'}$ to $\mathbb{R}_+^{R \times R'}$. For each $\theta \in [0, 1]^{R'}$, let $\widehat{P}(\theta)$ be the minimizer in (4.2) with $C(\theta)$ substituted for C . Then $\theta \mapsto \widehat{P}(\theta)$ is Lipschitz continuous from $[0, 1]^{R'}$ to $\mathbb{R}_+^{R \times R'}$.*

Indeed, $\theta \mapsto C(\mathbf{z}, \mathbf{z}'(\theta))$ and $\theta \mapsto C(\mathbf{z}'(\theta), \mathbf{z}'(\theta))$ (defined in Section 4.8.1.2) are obviously bounded and Lipschitz continuous.

Let us prove Lemma 10. By [Cuturi and Doucet, 2014, Proposition 4.3], for every $\theta \in \mathbb{R}^{R'}$,

$$\widehat{P}(\theta) = \text{diag}(\widehat{u}(\theta))K(\theta)\text{diag}(\widehat{v}(\theta)),$$

where $\widehat{u} : \mathbb{R}^{R'} \rightarrow (\mathbb{R}_+^*)^R$, $\widehat{v} : \mathbb{R}^{R'} \rightarrow (\mathbb{R}_+^*)^{R'}$ and the Gibbs kernel functions $K : \mathbb{R}^{R'} \rightarrow \mathbb{R}^{R \times R'}$, given by

$$K(\theta) := \left(\exp \left[- (C(\theta))_{r,r'} / \gamma \right] \right)_{r \in [R], r' \in [R']}$$

satisfy the mass conservation constraints inherent to $\Pi_{R,R'}$:

$$\text{diag}(\widehat{u}(\theta))K(\theta)\text{diag}(\widehat{v}(\theta))\mathbf{1}_{R'} = \frac{1}{R}\mathbf{1}_R \quad (4.28)$$

$$\text{diag}(\widehat{v}(\theta))K(\theta)^\top \text{diag}(\widehat{u}(\theta))\mathbf{1}_R = \frac{1}{R'}\mathbf{1}_{R'}, \quad (4.29)$$

Equivalently, using the entrywise division of vectors,

$$\widehat{u}(\theta) = \frac{\frac{1}{R}\mathbf{1}_R}{K(\theta)\widehat{v}(\theta)}, \quad \widehat{v}(\theta) = \frac{\frac{1}{R'}\mathbf{1}_{R'}}{K(\theta)^\top \widehat{u}(\theta)}. \quad (4.30)$$

Note that $(\rho\widehat{u}(\theta), \widehat{v}(\theta)/\rho)$ also satisfy (4.28) and (4.29) for any $\rho > 0$. Thus, without loss of generality, we can impose from now on that, for all $\theta \in \text{dom } g_\tau$, the first element $\widehat{u}_1(\theta)$ of $\widehat{u}(\theta)$ equals 1 (this affects both $\widehat{u}(\theta)$ and $\widehat{v}(\theta)$).

We now consider the following steps.

- The Gibbs kernel function K is Lipschitz continuous on $\text{dom } g_\tau$ with Lipschitz constant $L_K := k_u^2 L_C^2 / \gamma^2$ where $k_u := \max\{(K(\theta))_{r,r'} : r \in \llbracket R \rrbracket, r' \in \llbracket R' \rrbracket\}$ and L_C is the Lipschitz constant of $\theta \mapsto C(\theta)$.

Proof: The function $\theta \mapsto C(\theta)$ is bounded, so $\theta \mapsto K(\theta)$ is bounded as well. For all $\theta, \theta' \in [0, 1]^{R'}$, $r \in \llbracket R \rrbracket$ and $r' \in \llbracket R' \rrbracket$, it holds that

$$\begin{aligned} & |(K(\theta))_{r,r'} - (K(\theta'))_{r,r'}| \\ &= \max\{e^{-(C(\theta))_{r,r'}/\gamma}, e^{-(C(\theta'))_{r,r'}/\gamma}\} \times |1 - \exp(-|(C(\theta))_{r,r'} - (C(\theta'))_{r,r'}|/\gamma)| \\ &\leq \frac{k_u}{\gamma} \times |(C(\theta))_{r,r'} - (C(\theta'))_{r,r'}|. \end{aligned}$$

Therefore,

$$\begin{aligned} \|K(\theta) - K(\theta')\|_F^2 &= \sum_{r \in \llbracket R \rrbracket, r' \in \llbracket R' \rrbracket} [(K(\theta))_{r,r'} - (K(\theta'))_{r,r'}]^2 \\ &\leq \frac{k_u^2}{\gamma^2} \sum_{r \in \llbracket R \rrbracket, r' \in \llbracket R' \rrbracket} [(C(\theta))_{r,r'} - (C(\theta'))_{r,r'}]^2 \\ &\leq \frac{k_u^2 L_C^2}{\gamma^2} \|\theta - \theta'\|_2^2. \end{aligned}$$

- Denote $k_\ell := \min\{(K(\theta))_{r,r'} : r \in \llbracket R \rrbracket, r' \in \llbracket R' \rrbracket\}$. For every $\theta \in \text{dom } g_\tau$,

$$\lambda(K(\theta)) \leq \Lambda := (k_u - k_\ell)/(k_u + k_\ell) < 1. \quad (4.31)$$

Proof: Because $k_\ell \leq (K(\theta))_{r,r'} \leq k_u$ for all $\theta \in \text{dom } g_\tau$, $r \in \llbracket R \rrbracket$, $r' \in \llbracket R' \rrbracket$, it holds that $(K(\theta))_{i,k}(K(\theta))_{j,\ell}/((K(\theta))_{j,k}(K(\theta))_{i,\ell}) \leq k_u^2/k_\ell^2$ for all $i, j \in \llbracket R \rrbracket, k, \ell \in \llbracket R' \rrbracket$. Consequently, $\eta(K(\theta)) \leq k_u^2/k_\ell^2$ hence $\lambda(K(\theta)) = (\sqrt{\eta(K)} - 1)/(\sqrt{\eta(K)} + 1) \leq (k_u - k_\ell)/(k_u + k_\ell)$.

- For every $\theta \in \text{dom } g_\tau$, $\hat{u}(\theta)$ and $\hat{v}(\theta)$ are uniformly bounded: for all $r \in \llbracket R \rrbracket$, $r' \in \llbracket R' \rrbracket$,

$$\frac{k_\ell}{k_u R'} \leq \hat{u}_r(\theta) \leq \frac{k_u R}{k_\ell}, \quad (4.32)$$

$$\frac{k_\ell}{k_u^2 R' R^2} \leq \hat{v}_{r'}(\theta) \leq \frac{1}{k_\ell R}. \quad (4.33)$$

Proof: Set arbitrarily $\theta \in \text{dom } g_\tau$. In view of (4.28) (first row), since $\hat{u}_1(\theta) = 1$, we have

$$k_\ell \|\hat{v}(\theta)\|_\infty \leq \frac{1}{R} = \sum_{r' \in \llbracket R' \rrbracket} (K(\theta))_{1r'} \hat{v}_{r'}(\theta) \leq k_u R' \|\hat{v}(\theta)\|_\infty. \quad (4.34)$$

Set $r'_0 \in \arg \max\{\hat{v}_i(\theta) : i \in \llbracket R' \rrbracket\}$. In view of (4.29) (r' th row), we have

$$\frac{1}{R'} = \hat{v}_{r'_0}(\theta) \sum_{r \in \llbracket R \rrbracket} (K(\theta))_{rr'_0} \hat{u}_r(\theta) \geq k_\ell \|\hat{v}(\theta)\|_\infty \|\hat{u}(\theta)\|_\infty.$$

Hence, by (4.34),

$$\|\widehat{u}(\theta)\|_\infty \leq \frac{1}{k_\ell R' \|\widehat{v}(\theta)\|_\infty} \leq \frac{k_u R R'}{k_\ell R'} = \frac{k_u R}{k_\ell}. \quad (4.35)$$

Furthermore, for any $r' \in \llbracket R' \rrbracket$, in view of (4.29) (r' th row) and (4.35),

$$\frac{1}{R'} = \widehat{v}_{r'}(\theta) \sum_{r \in \llbracket R \rrbracket} (K(\theta))_{rr'} \widehat{u}_r(\theta) \leq R k_u \|\widehat{u}(\theta)\|_\infty \widehat{v}_{r'}(\theta) \leq \frac{k_u^2 R^2}{k_\ell} \widehat{v}_{r'}(\theta). \quad (4.36)$$

The inequalities (4.34) and (4.36) readily imply (4.33). Likewise, for any $r \in \llbracket R \rrbracket$, in view of (4.28) (r th row),

$$\frac{1}{R} = \widehat{u}_r(\theta) \sum_{r' \in \llbracket R' \rrbracket} (K(\theta))_{rr'} \widehat{v}_{r'}(\theta) \leq R' k_u \|\widehat{v}(\theta)\|_\infty \widehat{u}_r(\theta) \leq \frac{k_u R'}{k_\ell R} \widehat{u}_r(\theta). \quad (4.37)$$

The inequalities (4.35) and (4.37) readily imply (4.32).

— The function $\theta \mapsto \widehat{u}(\theta)$ is Lipschitz continuous on $\text{dom } g_\tau$ with Lipschitz constant

$$L_{\widehat{u}} := \frac{2k_u^3 R^2 \sqrt{R'} L_K}{(1 - \Lambda^2) k_\ell^4} (\sqrt{R} + \Lambda \sqrt{R'}).$$

Proof. Set arbitrarily $\theta, \theta' \in \text{dom } g_\tau$. Inequalities (4.32) and (4.33) imply that

$$\begin{aligned} \min\{(K(\theta)\widehat{v}(\theta'))_r : r \in \llbracket R \rrbracket\} &\geq k_\ell^2 / (k_u^2 R^2), \\ \min\{(K(\theta)^\top \widehat{u}(\theta'))_{r'} : r' \in \llbracket R' \rrbracket\} &\geq k_\ell^2 R / (k_u R'). \end{aligned}$$

In view of Lemma 8 (first inequality), (4.21) (second inequality), (4.33) and the fact that K is L_K -Lipschitz (third inequality), we obtain

$$\begin{aligned} d_{\mathcal{H}}(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta)) &\leq \frac{2k_u^2 R^2}{k_\ell^2} \|K(\theta)\widehat{v}(\theta) - K(\theta')\widehat{v}(\theta)\|_2 \\ &\leq \frac{2k_u^2 R^2}{k_\ell^2} \|K(\theta) - K(\theta')\|_F \|\widehat{v}(\theta)\|_2 \\ &\leq \frac{2k_u^2 R \sqrt{R'} L_K}{k_\ell^3} \|\theta - \theta'\|_2. \end{aligned} \quad (4.38)$$

Likewise, using (4.32) instead of (4.33)

$$\begin{aligned} d_{\mathcal{H}}(K(\theta)^\top \widehat{u}(\theta), K(\theta')^\top \widehat{u}(\theta)) &\leq \frac{2k_u R'}{k_\ell^2 R} \|K(\theta_1)^\top \widehat{u}(\theta_1) - K(\theta_2)^\top \widehat{u}(\theta_1)\|_2 \\ &\leq \frac{2k_u R'}{k_\ell^2 R} \|K(\theta)^\top - K(\theta')^\top\|_F \|\widehat{u}(\theta)\|_2 \\ &\leq \frac{2k_u^2 \sqrt{R'} L_K}{k_\ell^3} \|\theta - \theta'\|_2. \end{aligned} \quad (4.39)$$

We can now bound the Hilbert projective metric between $\widehat{v}(\theta)$ and $\widehat{v}(\theta')$: by invoking in turn (4.30), (4.26), the triangle inequality, (4.27) and both (4.39) and (4.31), we get

$$\begin{aligned}
d_{\mathcal{H}}(\widehat{v}(\theta), \widehat{v}(\theta')) &= d_{\mathcal{H}}\left(\frac{\mathbf{1}_{R'}/R'}{K(\theta)^\top \widehat{u}(\theta)}, \frac{\mathbf{1}_{R'}/R'}{K(\theta')^\top \widehat{u}(\theta')}\right) \\
&= d_{\mathcal{H}}\left(K(\theta)^\top \widehat{u}(\theta), K(\theta')^\top \widehat{u}(\theta')\right) \\
&\leq d_{\mathcal{H}}\left(K(\theta)^\top \widehat{u}(\theta), K(\theta')^\top \widehat{u}(\theta)\right) + d_{\mathcal{H}}\left(K(\theta')^\top \widehat{u}(\theta), K(\theta')^\top \widehat{u}(\theta')\right) \\
&\leq d_{\mathcal{H}}\left(K(\theta)^\top \widehat{u}(\theta), K(\theta')^\top \widehat{u}(\theta)\right) + \lambda(K(\theta')) d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta')) \\
&\leq \frac{2k_u^2 \sqrt{R} R' L_K}{k_\ell^3} \|\theta - \theta'\|_2 + \Lambda d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta')). \tag{4.40}
\end{aligned}$$

Likewise, by invoking in turn (4.30), (4.26), the triangle inequality, (4.27) and both (4.39) and (4.40), we get

$$\begin{aligned}
d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta')) &= d_{\mathcal{H}}\left(\frac{\mathbf{1}_R/R}{K(\theta)\widehat{v}(\theta)}, \frac{\mathbf{1}_R/R}{K(\theta')\widehat{v}(\theta')}\right) \\
&= d_{\mathcal{H}}\left(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta')\right) \\
&\leq d_{\mathcal{H}}\left(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta)\right) + d_{\mathcal{H}}\left(K(\theta')\widehat{v}(\theta), K(\theta')\widehat{v}(\theta')\right) \\
&\leq d_{\mathcal{H}}\left(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta)\right) + \lambda(K(\theta')) d_{\mathcal{H}}(\widehat{v}(\theta), \widehat{v}(\theta')) \\
&\leq d_{\mathcal{H}}\left(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta)\right) \\
&\quad + \Lambda \left(\frac{2k_u^2 \sqrt{R} R' L_K}{k_\ell^3} \|\theta - \theta'\|_2 + \Lambda d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta')) \right).
\end{aligned}$$

The above inequality and (4.38) then yield

$$\begin{aligned}
d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta')) &\leq \frac{1}{1 - \Lambda^2} \left(d_{\mathcal{H}}\left(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta)\right) + \Lambda \frac{2k_u^2 \sqrt{R} R' L_K}{k_\ell^3} \|\theta - \theta'\|_2 \right) \\
&\leq \frac{2k_u^2 \sqrt{R} R' L_K}{(1 - \Lambda^2) k_\ell^3} (\sqrt{R} + \Lambda \sqrt{R'}) \|\theta - \theta'\|_2.
\end{aligned}$$

Therefore, by Lemma 8 and (4.32), $\|\widehat{u}(\theta) - \widehat{u}(\theta')\|_2 \leq L_{\widehat{u}} \|\theta - \theta'\|_2$, which completes the proof.

— The function $\theta \mapsto \widehat{v}(\theta)$ is Lipschitz continuous on $\text{dom } g_\tau$ with Lipschitz constant

$$L_{\widehat{v}} := \frac{k_u L_K}{k_\ell^3 \sqrt{R}} + \frac{k_u \sqrt{R'} L_{\widehat{u}}}{k_\ell^2 R^{3/2}}.$$

Proof: Set arbitrarily $\theta, \theta' \in \text{dom } g_\tau$. By (4.30) and (4.33),

$$\|\widehat{v}(\theta) - \widehat{v}(\theta')\|_2 = \left\| \frac{\mathbf{1}_{R'}/R'}{K(\theta)^\top \widehat{u}(\theta)} - \frac{\mathbf{1}_{R'}/R'}{K(\theta')^\top \widehat{u}(\theta')} \right\|_2$$

$$\begin{aligned}
&\leq \frac{\|K(\theta)^\top \hat{u}(\theta) - K(\theta')^\top \hat{u}(\theta')\|_2}{R' \min_{r' \in \llbracket R' \rrbracket} \{(K(\theta_1)^\top \hat{u}(\theta_1))_{r'}\} \min_{r' \in \llbracket R' \rrbracket} \{(K(\theta')^\top \hat{u}(\theta'))_{r'}\}} \\
&= \frac{\|K(\theta)^\top \hat{u}(\theta) - K(\theta')^\top \hat{u}(\theta')\|_2}{\min_{r' \in \llbracket R' \rrbracket} \{\hat{v}_{r'}(\theta)^{-1}\} \min_{r' \in \llbracket R' \rrbracket} \{\hat{v}_{r'}(\theta')^{-1}\}} \\
&\leq \frac{1}{k_\ell^2 R^2} \|K(\theta)^\top \hat{u}(\theta) - K(\theta')^\top \hat{u}(\theta')\|_2.
\end{aligned}$$

Moreover, using in turn the triangle inequality, (4.21) then the fact that K and \hat{u} are Lipschitz continuous and bounded on $\text{dom } g_\tau$, we get

$$\begin{aligned}
\|K(\theta)^\top \hat{u}(\theta) - K(\theta')^\top \hat{u}(\theta')\|_2 &\leq \|K(\theta)^\top \hat{u}(\theta) - K(\theta')^\top \hat{u}(\theta)\|_2 \\
&\quad + \|K(\theta')^\top \hat{u}(\theta) - K(\theta')^\top \hat{u}(\theta')\|_2 \\
&\leq \|K(\theta) - K(\theta')\|_F \|\hat{u}(\theta)\|_2 \\
&\quad + \|K(\theta')\|_F \|\hat{u}(\theta) - \hat{u}(\theta')\|_2 \\
&\leq \left(\frac{k_u R^{3/2} L_K}{k_\ell} + \sqrt{RR'} k_u L_{\hat{u}} \right) \|\theta - \theta'\|_2.
\end{aligned}$$

Therefore, $\|\hat{v}(\theta) - \hat{v}(\theta')\|_2 \leq L_{\hat{v}} \|\theta - \theta'\|_2$, which completes the proof.

— The function $\hat{P}(\theta)$ is Lipschitz continuous on $\text{dom } g_\tau$.

Proof: We have proved that $\theta \mapsto \hat{u}$, $\theta \mapsto K(\theta)$ and $\theta \mapsto \hat{v}(\theta)$ are bounded and Lipschitz continuous on $\text{dom } g_\tau$. Consequently, so is

$$\theta \mapsto \hat{P}(\theta) = \text{diag}(\hat{u}(\theta)) K(\theta) \text{diag}(\hat{v}(\theta)).$$

This completes the proof of Lemma 10, hence that of the fact that $\theta \mapsto \hat{P}_\theta$ and $\theta \mapsto \hat{Q}_\theta$ are Lipschitz continuous on $\text{dom } g_\tau$.

4.8.1.4 The gradient of f is Lipschitz continuous

Set arbitrarily $\theta, \theta' \in \text{dom } g_\tau \subset [0, 1]^N$. We begin by noting that, by the triangle inequality and (4.21),

$$\begin{aligned}
\frac{1}{2} \|\nabla f(\theta) - \nabla f(\theta')\|_2 &\leq \|y\|_2 \times \|\hat{P}_\theta - \hat{P}_{\theta'}\|_F + \|\theta\|_2 \times \|\hat{Q}_\theta - \hat{Q}_{\theta'}\|_F + \|\hat{Q}_{\theta'}\|_F \times \|\theta - \theta'\|_2 \\
&\leq \|y\|_2 \times \|\hat{P}_\theta - \hat{P}_{\theta'}\|_F + \sqrt{N} \times \|\hat{Q}_\theta - \hat{Q}_{\theta'}\|_F + \|\theta - \theta'\|_2.
\end{aligned}$$

We then readily conclude because we showed in Section 4.8.1.3 that $\theta \mapsto \hat{P}_\theta$ and $\theta \mapsto \hat{Q}_\theta$ are Lipschitz continuous on $\text{dom } g_\tau$.

4.8.2 The function H_δ satisfies the Kurdyka-Lojasiewicz property

4.8.2.1 The Kurdyka-Lojasiewicz property

Let us first recall what is the Kurdyka-Lojasiewicz property. Let $\ell : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous function. For any $-\infty < \eta_1 < \eta_2 \leq +\infty$, the bracket

$[\eta_1 < \ell < \eta_2]$ is the set $\{x \in \mathbb{R}^d : \eta_1 < \ell(x) < \eta_2\}$. We refer the reader to [Attouch et al., 2010, Section 2] for elementary facts of nonsmooth analysis, including the definition of $\partial\ell$, the limiting-subdifferential of ℓ [Rockafellar and Wets, 1998].

Définition 4 (Kurdyka-Lojasiewicz property, definition 3.1 in Attouch et al. [2010]). *The function ℓ is said to have the Kurdyka-Lojasiewicz property at $\bar{x} \in \text{dom } \partial\ell$ if there exists $\eta \in (0, +\infty]$, a neighborhood U of \bar{x} and a continuous concave function $\varphi : [0, \eta) \rightarrow \mathbb{R}_+$ such that:*

- $\varphi(0) = 0$,
- φ is C^1 on $(0, \eta)$,
- for all $s \in (0, \eta)$, $\varphi'(s) > 0$,
- and for all $x \in U \cap [\ell(\bar{x}) < \ell < \ell(\bar{x}) + \eta]$, the Kurdyka-Lojasiewicz inequality holds:

$$\varphi'(\ell(x) - \ell(\bar{x})) \text{dist}(0, \partial\ell(x)) \geq 1. \quad (4.41)$$

Inequality (4.41) can be interpreted as follows: subject to the reparametrization of f through φ , we deal with a sharp function. To see this, consider the simple case where the finite-valued f is differentiable and $f(\bar{x}) = 0$, so that (4.41) rewrites as $\|\nabla\varphi \circ f(x)\| \geq 1$: the function φ transforms a singular region, characterized by arbitrarily small gradients, into a regular region where the gradients are bounded away from zero. Thus the transformation φ is aptly referred to as a “desingularizing function” for f . For further theoretical and geometrical insights, we refer to [Bolte et al., 2010].

To prove that H_δ satisfies the Kurdyka-Lojasiewicz property, we apply Theorem 4.1 in [Attouch et al., 2010]. We state it below for the sake of completeness. The key notions necessary to understand the theorem are succinctly presented after the statement.

Theorem 11 (Theorem 4.1 in Attouch et al. [2010]). *Any proper lower semicontinuous function $\ell : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ which is definable in an o -minimal structure \mathcal{O} over \mathbb{R} has the Kurdyka-Lojasiewicz property at each point of $\text{dom } \partial\ell$. Moreover the function φ appearing in (4.41) is definable in \mathcal{O} .*

4.8.2.1.1 On o -minimal structures. An o -minimal structure over \mathbb{R} can be viewed as an axiomatization of the quantitative properties of semialgebraic sets. Semialgebraic sets are finite unions and intersections of sets of the form $\{x \in \mathbb{R}^d : Q(x) = 0, R(x) < 0\}$ for some polynomial functions $Q, R : \mathbb{R}^d \rightarrow \mathbb{R}$. Algebraic sets are finite unions and intersections of sets of the form $\{x \in \mathbb{R}^d : Q(x) = 0\}$ for some polynomial function $Q : \mathbb{R}^d \rightarrow \mathbb{R}$.

Formally, a collection $\mathcal{O} = \{\mathcal{O}_n\}_{n \geq 0}$ is a structure over \mathbb{R} if the following conditions are met:

- (a) for each $n \geq 0$, \mathcal{O}_n is a collection of subsets of \mathbb{R}^n ;
- (b) for each $n \geq 0$, all algebraic subsets of \mathbb{R}^n are in \mathcal{O}_n ;
- (c) for each $n \geq 0$, \mathcal{O}_n is a Boolean subalgebra, that is, $\emptyset \in \mathcal{O}_n$ and, for every $A, B \in \mathcal{O}_n$, $A \cup B$, $A \cap B$ and $\mathbb{R}^n \setminus A$ belong to \mathcal{O}_n ;

- (d) if $A \in \mathcal{O}_m$ and $B \in \mathcal{O}_n$, then $A \times B \in \mathcal{O}_{m+n}$;
- (e) if $p : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$ is the projection on the first n coordinates and $A \in \mathcal{O}_{n+1}$, then $p(A) \in \mathcal{O}_n$.

It is o -minimal if, in addition,

- (f) the elements of \mathcal{O}_1 are precisely the finite unions of intervals.

The smallest o -minimal structure over \mathbb{R} containing the semialgebraic sets is denoted \mathbb{R}_{alg} . It is the collection $\{\mathcal{O}_n\}_{n \geq 0}$ where each \mathcal{O}_n is the class of semialgebraic sets on \mathbb{R}^n [Benedetti and Risler, 1990, Bochnak et al., 1998].

The smallest structure containing the semialgebraic sets and the graph of the exponential function $\exp : \mathbb{R} \rightarrow \mathbb{R}_+^*$ is denoted \mathbb{R}_{exp} . It extends \mathbb{R}_{alg} and it is o -minimal over \mathbb{R} [Wilkie, 1996].

4.8.2.1.2 On definable sets and definable functions. Given an o -minimal structure $\mathcal{O} = (\mathcal{O}_n)_{n \geq 0}$ over \mathbb{R} , the elements of each \mathcal{O}_n are called the definable subsets of \mathbb{R}^n . A function $\varphi : A \rightarrow B$ between two definable sets is definable in \mathcal{O} if its graph is definable in \mathcal{O} .

For instance, a polynomial function $Q : \mathbb{R}^d \rightarrow \mathbb{R}$ is definable in \mathbb{R}_{alg} , hence in \mathbb{R}_{exp} as well.

We use the following properties [Attouch et al., 2010] (from now on, we write “definable” in lieu of “definable in \mathcal{O} ”):

- (g) if $\varphi : A \rightarrow B$ is definable and if $A' \subset A$ is definable, then $\varphi|_{A'}$ is definable;
- (h) if φ is definable, then $|\varphi|$ is definable;
- (i) finite sums of definable function are definable;
- (j) any indicator function $\mathbf{I}\{A\}$ (which equals 0 if the argument falls in A and $+\infty$ otherwise) of a definable set A is definable;
- (k) generalized inverse functions of definable functions are definable;
- (l) compositions of definable functions are definable;
- (m) if ψ and C are definable, then $\mathbb{R}^n \ni x \mapsto \inf_{y \in C} \psi(x, y)$ and $\mathbb{R}^n \ni x \mapsto \sup_{y \in C} \psi(x, y)$ are definable.

4.8.2.2 The function H_δ is definable in \mathbb{R}_{exp}

Let us prove now that H_δ is definable in \mathbb{R}_{exp} – from now on, “definable” means definable in \mathbb{R}_{exp} . We consider the following steps.

— The set $\Pi_{R, R'}$ is semialgebraic hence definable.

Proof: Introduce the sets $A_{r, r'} := \{P \in \mathbb{R}^{R \times R'} : P_{r, r'} \geq 0\}$, $B_r := \{P \in \mathbb{R}^{R \times R'} : \sum_{r' \in \llbracket R' \rrbracket} P_{r, r'} = \frac{1}{R}\}$ and $C_{r'} := \{P \in \mathbb{R}^{R \times R'} : \sum_{r \in \llbracket R \rrbracket} P_{r, r'} = \frac{1}{R'}\}$ (for all $r \in \llbracket R \rrbracket$ and $r' \in \llbracket R' \rrbracket$). Each of them is semialgebraic. Therefore their intersection, which equals $\Pi_{R, R'}$, is semialgebraic too, hence definable.

— Consider $F : \mathbb{R}^N \times \mathbb{R}^{M \times N} \times \mathbb{R}^{N \times N} \rightarrow \mathbb{R}$ given by

$$F(\theta, P, Q) := \sum_{m \in \llbracket M \rrbracket, n \in \llbracket N \rrbracket} P_{m,n} \left(d(x_m, x'_n)^2 + (y_m - \theta_n)^2 \right) - \frac{1}{2} \sum_{m \in \llbracket M \rrbracket, n \in \llbracket N \rrbracket} Q_{n,n'} \left(d(x'_n, x'_{n'})^2 + (\theta_n - \theta_{n'})^2 \right) + g_\tau(\theta).$$

Proof: The function $(\theta, P) \mapsto F(\theta, P, Q) - g_\tau(\theta)$ is definable because it is polynomial. Moreover, g_τ is also definable.

— When $g_\tau(\theta) = \tau \|\theta\|_1 + \mathbf{I}\{\theta \in [0, 1]^N\}$: on the one hand, $\theta \mapsto \|\theta\|_1 = \sum_{n \in \llbracket N \rrbracket} |\theta_n|$ is definable as a finite sum of definable functions (properties (i) and (h)); on the other hand, $\mathbf{I}\{[0, 1]^N\}$ is definable because $[0, 1]^N$ is definable (property (j)). Therefore, g_τ is definable (property (i)).

— When $g_\tau(\theta) = \mathbf{I}\{\|\theta\|_1 \leq \tau\} + \mathbf{I}\{\theta \in [0, 1]^N\}$: on the one hand, the set $\{\theta \in \mathbb{R}^N : \|\theta\|_1 \leq \tau\}$ is definable because it can be written as

$$\bigcup_{\varepsilon \in \{\pm 1\}^N} \left[\bigcap_{n \in \llbracket N \rrbracket} \{\theta \in \mathbb{R}^N : \varepsilon_n \theta_n \geq 0\} \cap \{\theta \in \mathbb{R}^N : \sum_{n \in \llbracket N \rrbracket} \varepsilon_n \theta_n - \tau \leq 0\} \right],$$

which is semialgebraic since it is a finite union and intersection of semialgebraic sets; therefore, $\theta \mapsto \mathbf{I}\{\|\theta\|_1 \leq \tau\}$ is definable (property (j)). On the other hand, we already proved that $\mathbf{I}\{[0, 1]^N\}$ is definable, hence g_τ is definable (property (i)).

It follows that F is definable (property (i)). Because the set $\mathbb{R}^N \times \Pi_{M,N} \times \Pi_{N,N}$ is definable, this implies that $F|_{\mathbb{R}^N \times \Pi_{M,N} \times \Pi_{N,N}}$ is definable (property (g)).

— The function $\gamma E : P \mapsto \gamma \times E(P)$ from $\Pi_{R,R'}$ to \mathbb{R} is definable.

Proof: The function $\log : \mathbb{R}_+^* \rightarrow \mathbb{R}$ is definable (property (k)). Consequently, $\varphi : \mathbb{R}_+^* \rightarrow \mathbb{R}^2$ given by $\varphi(x) := (\log(x), x)$ is definable because its graph can be written as

$$(\Gamma_{\log} \times \mathbb{R}) \cap \{(x, y, z) \in \mathbb{R}^3 : x - z = 0\}$$

where the graph Γ_{\log} of \log is definable and the right-hand-side set is algebraic hence definable, revealing that the graph of φ is definable as the intersection of two definable sets. Moreover, the polynomial function $Q : \mathbb{R}^2 \rightarrow \mathbb{R}$ given by $Q(x, y) := -\gamma x(y - 1)$ is definable. Therefore, $\phi := Q \circ \varphi : \mathbb{R}_+^* \rightarrow \mathbb{R}$, so that $\phi(x) = -\gamma x(\log(x) - 1)$, is definable (property (l)). Setting $\phi(0) := 0$ extends ϕ by continuity and yields a definable function $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}$. It follows that $\gamma \mathcal{E} : (\mathbb{R}_+)^{R \times R'} \rightarrow \mathbb{R}$ given by $\gamma \mathcal{E}(P) := \sum_{r \in \llbracket R \rrbracket, r' \in \llbracket R' \rrbracket} \phi(P_{r,r'})$ is definable (property (i)), hence $\gamma E := \gamma \mathcal{E}|_{\Pi_{R,R'}}$ is definable too (property (g)).

— The function $(f + g_\tau) : \mathbb{R}^N \rightarrow \mathbb{R}$ is definable.

Proof: This is a straightforward consequence of the fact that, for all $\theta \in \mathbb{R}^N$,

$$(f + g_\tau)(\theta) := \min_{P \in \Pi_{M,N}} \max_{Q \in \Pi_{N,N}} \left\{ F|_{\mathbb{R}^N \times \Pi_{M,N} \times \Pi_{N,N}} + \gamma E(P) - \frac{1}{2} \gamma E(Q) \right\},$$

where the sets $\Pi_{M,N}$ and $\Pi_{N,N}$ are definable (property (m)).

— The function H_δ is definable.

Proof: Recall that $H_\delta : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ is given by $H_\delta(\theta, \theta') := f(\theta') + g_\tau(\theta') + \delta\|\theta - \theta'\|_2^2$. The function $(\theta, \theta') \mapsto f(\theta') + g_\tau(\theta')$ between $\mathbb{R}^N \times \mathbb{R}^N$ and \mathbb{R} is definable because its graph

$$\{(\theta, \theta', f(\theta') + g_\tau(\theta')) : (\theta, \theta') \in \mathbb{R}^N \times \mathbb{R}^N\} = \mathbb{R}^N \times \Gamma_{f+g_\tau},$$

where Γ_{f+g_τ} is the graph of $(f + g_\tau)$, is definable as the product of two definable sets. Moreover, the function $(\theta, \theta') \mapsto \delta\|\theta - \theta'\|_2^2$ between $\mathbb{R}^N \times \mathbb{R}^N$ and \mathbb{R} is polynomial, hence definable. Therefore, H_δ is definable (property (i)).

4.8.2.3 The function H_δ is proper and lower semicontinuous, hence satisfies the Kurdyka-Lojasiewicz property on the domain of ∂H_δ

The function H_δ never takes on the value $-\infty$ and $H_\delta(0)$ is finite, so H_δ is proper. Moreover, f is differentiable (see Section 4.8.1), g_τ is lower semicontinuous because it is either continuous (when $g_\tau(\cdot) = \tau\|\cdot\|_1$) or lower semicontinuous (when g_τ is the characteristic function of the closed $\|\cdot\|_1$ -ball centered at 0 and with radius τ), and $(\theta, \theta') \mapsto \delta\|\theta - \theta'\|_2^2$ is continuous. Therefore, H_δ is proper and lower semicontinuous. By Theorem 11, H_δ satisfies the Kurdyka-Lojasiewicz property on the domain of ∂H_δ .

Chapitre 5

Conclusion et perspectives

Les travaux menés dans cette étude ont mis en exergue les apports du machine learning pour l'estimation des dommages liés à la survenance d'un événement de sécheresse RGA. Ce phénomène a été modélisé comme une série temporelle courte, constituée à chaque pas de temps de nombreuses observations faiblement dépendantes. Inspiré du Super Learner, un algorithme d'agrégation de modèles, un algorithme original a ensuite été développé. L'overarching Super Learner combine les prédictions de plusieurs meta-algorithmes, chaque meta-algorithme combinant lui-même les prédictions d'algorithmes fondamentaux. L'analyse théorique repose notamment sur une hypothèse de stationnarité du mécanisme produisant les dommages à partir des covariables, tant dans l'espace que dans le temps. Elle repose également sur la modélisation de la dépendance entre les communes via un graphe. Sous ces conditions, une inégalité oracle a été proposée et constitue une garantie théorique d'apprentissage pour l'algorithme original développé. Dans la pratique, les résultats obtenus pour l'estimation des dommages sont tout à fait satisfaisants et constituent une amélioration par rapport au modèle actuel dont dispose CCR. En revanche, la transposition brutale de cette démarche à l'estimation des probabilités de demande de reconnaissance, une étape intermédiaire pour l'estimation des dommages, ne permet pas d'obtenir des performances intéressantes. La prise en compte d'une nouvelle source de données, le recours à une modélisation alternative du problème et l'emploi du transfer learning ont permis d'améliorer significativement les performances de l'overarching Super Learner pour la prédiction des probabilités de demande de reconnaissance. Enfin, un nouvel algorithme a été proposé dans cette étude pour l'estimation de ces probabilités. Celui-ci est qualifié d'hybride dans la mesure où il repose à la fois sur un Super Learner et sur le transport optimal.

Si les performances de cet algorithme hybride ne constituent pas une amélioration par rapport à celles obtenues par l'overarching Super Learner dans le cadre de l'approche dynamique avec transfer learning, il existe toutefois un potentiel de synergie important entre ces deux approches. Il serait par exemple intéressant d'intégrer l'algorithme hybride dans la collection d'algorithmes fondamentaux de l'overarching Super Learner afin d'augmenter la diversité des modèles représentés, et donc potentiellement les performances.

L'utilisation des prédictions de l'overarching Super Learner avec transfer learning dans la procédure hybride, plutôt que celles d'un Super Learner discret, représente un autre moyen de combiner les deux approches.

En outre, les liens entre les deux sous-problèmes que sont l'estimation des dommages pour les communes reconnues d'une part, et l'anticipation des communes demanderessees d'autre part pourraient faire l'objet d'une étude. En premier lieu, l'approche dynamique déployée pour l'estimation des probabilités de demande de reconnaissance pourrait être mise en œuvre pour l'estimation des dommages. Notamment, les fichiers communiqués à CCR par la commission interministérielle contiennent une indication du nombre de bâtiments touchés. Cette information s'avèrerait sans nul doute pertinente pour l'estimation du coût engendré par l'événement. Également, nous n'avons pas observé les performances que nous obtiendrions si nous réalisions l'estimation des dommages sur la base des probabilités de demande de reconnaissance estimées. Enfin, la tâche de prédiction des dommages et la tâche de prédiction des probabilités de demande de reconnaissance ont été réalisées de façon indépendante. Le multi-task learning permettrait de faire interagir les apprentissages des deux tâches, et par ce biais d'améliorer les performances obtenues pour chacune d'elles.

Par ailleurs, la prise en compte du changement climatique constitue également une perspective de recherche. Le transfer learning, ne reposant pas sur une hypothèse d'identique distribution entre les domaines source et cible, pourrait alors être mis à profit. En considérant que le changement climatique correspond à une dérive de la distribution du SWI pouvant aller jusqu'à une évolution du support de sa distribution, le transfer learning hétérogène constituerait alors une réponse adaptée. Ces travaux permettraient d'étudier l'impact du changement climatique sur la sinistralité sécheresse.

Finalement, la mise à disposition de données permettant de mieux caractériser l'aléa d'un événement sécheresse constitue un levier important pour l'amélioration des prédictions.

Bibliographie

- J. J. Allaire and F. Chollet. *keras : R Interface to 'Keras'*, 2021. URL <https://CRAN.R-project.org/package=keras>. R package version 2.4.0.
- Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7) :1545–1588, 1997.
- K. Antunez. *COGugaison :*, 2022. URL <https://antuki.github.io/COGugaison/>. R package version 1.0.5.
- J. Ardon. *Modélisation probabiliste de la dépendance spatiale et temporelle appliquée à l'étude du péril sécheresse dans le cadre du régime français d'indemnisation des catastrophes naturelles*. PhD thesis, Université de La Rochelle, 2014.
- H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems : an approach based on the Kurdyka-Lojasiewicz inequality. *Math. Oper. Res.*, 35(2) :438–457, 2010.
- Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4) :467–493, 2000.
- P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4) :1497–1537, 2005.
- P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101(473) :138–156, 2006.
- R. Benedetti and J.-J. Risler. *Real algebraic and semi-algebraic sets*. Actualités Mathématiques. [Current Mathematical Topics]. Hermann, Paris, 1990.
- D. Benkeser, C. Ju, S. Lendle, and M. J. van der Laan. Online cross-validation-based ensemble learning. *Stat. Med.*, 37(2) :249–260, 2018.
- P. N. Bennett. Using asymmetric distributions to improve text classifier probability estimates. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003. URL <https://api.semanticscholar.org/CorpusID:2382902>.

- B. Bercu, B. Delyon, and E. Rio. *Concentration inequalities for sums and martingales*. SpringerBriefs in Mathematics. Springer, Cham, 2015.
- D. P. Bertsekas. *Nonlinear programming*. Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont, MA, 1999.
- G. Birkhoff. Extensions of Jentzsch's theorem. *Trans. Amer. Math. Soc.*, 85, 1957.
- J. Bochnak, M. Coste, and M.-F. Roy. *Real algebraic geometry*, volume 36 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1998.
- J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of Lojasiewicz inequalities : Subgradient flows, talweg, convexity. *Trans. Amer. Math. Soc.*, 362 :3319–3363, 2010.
- S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities : A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013. ISBN 9780199535255. doi : 10.1093/acprof:oso/9780199535255.001.0001. URL <https://doi.org/10.1093/acprof:oso/9780199535255.001.0001>.
- R. B. Bradford. *Drought Events in Europe*, pages 7–20. Springer Netherlands, Dordrecht, 2000.
- L. Breiman. Bagging predictors. *Machine learning*, 24(2) :123–140, 1996a.
- L. Breiman. Stacked regressions. *Machine learning*, 24(1) :49–64, 1996b.
- L. Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.
- L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software : experiences from the scikit-learn project. In *ECML PKDD Workshop : Languages for Data Mining and Machine Learning*, pages 108–122, 2013.
- CCR. Modélisation de l'impact du changement climatique sur les dommages assurés dans le cadre du régime catastrophes naturelles. Technical report, Caisse Centrale de Réassurance, 2015. URL <https://www.ccr.fr/documents/35794/35836/Etude+climat.pdf/18d0afb3-0a2c-40a7-a5ca-8a10c570168e?t=1455202610000>.
- CCR. Conséquences du changement climatique sur le coût des catastrophes naturelles en France à l'horizon 2050. Technical report, Caisse Centrale de Réassurance, 2018. URL <https://www.ccr.fr/documents/35794/35836/Etude+Climatique+2018+version+complete.pdf/6a7b6120-7050-ff2e-4aa9-89e80c1e30f2?t=1536662736000#:~:text=A%20l%27horizon%202050%2C%20les,zones%20C3%A0%20risques%20pour%2015%25>.

-
- CCR. Les catastrophes naturelles en France : bilan 1982-2020. Technical report, Caisse Centrale de Réassurance, 2021. URL <https://side.developpement-durable.gouv.fr/ACCIDR/doc/SYRACUSE/795441>.
- CCR. Arrêtés de catastrophes naturelles. Technical report, Caisse Centrale de Réassurance, 2022a. URL <http://catastrophes-naturelles.ccr.fr/les-arretes>.
- CCR. Rapport d'activité 2021. Technical report, Caisse Centrale de Réassurance, 2022b. URL <https://www.ccr.fr/documents/35794/35839/CCR+RA+2021+web+all+24032022.pdf/84e4c7da-34b5-22e0-e048-06a0836b7392?t=1648135815072>.
- CCR. Les catastrophes naturelles en France : bilan 1982-2022. Technical report, Caisse Centrale de Réassurance, 2023.
- N. Cesa-Bianchi and C. Gentile. Improved risk tail bounds for on-line algorithms. *IEEE Trans. Inform. Theory*, 54(1) :386–390, 2008.
- N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge, 2006.
- A. Chambaz and G. Ecoto. *SequentialSuperLearner : sequential Super Learner Prediction*, 2021. URL <https://github.com/achambaz/SequentialSuperLearner>. R package version 0.0.0.9000.
- A. Charpentier, M. James, and H. Ali. Predicting drought and subsidence risks in France. *Nat. Hazards Earth Syst. Sci.*, 22 :2401–2418, 2022. doi : 10.5194/nhess-22-2401-2022.
- P. Chatelain and S. Loisel. Subsidence and household insurances in France : geolocated data and insurability. Technical report, 2021. URL <https://hal.science/hal-03791154>.
- T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, and Y. Li. *xgboost : Extreme Gradient Boosting*, 2021. URL <https://CRAN.R-project.org/package=xgboost>. R package version 1.4.1.1.
- M. Cuturi. Sinkhorn distances : Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693. PMLR, 22–24 Jun 2014.
- J. M. Danskin. The theory of max – min, with applications. *SIAM J. Appl. Math.*, 14, 1966.

- J. Dedecker. Exponential inequalities and functional central limit theorems for a random fields. *ESAIM Probab. Statist.*, 5 :77–104, 2001.
- M. Denuit and A. Charpentier. *Mathématiques de l'assurance non-vie*. 01 2004.
- B. Devkota, M. R. Karim, M. M. Rahman, and H. B. K. Nguyen. Accounting for expansive soil movement in geotechnical design : A state-of-the-art review. *Sustainability*, 14 :15662, 2022.
- P. A. Dirmeyer, A. J. Dolman, and N. Sato. The pilot phase of the global soil wetness project. *Bulletin of the American Meteorological Society*, 80(5) :851–878, 1999.
- P. Doukhan, J. León, and F. Portal. Vitesse de convergence dans le théorème central limite pour des variables aléatoires mélangeantes à valeurs dans un espace de Hilbert. *C. R. Acad. Sci. Paris Sér. I Math.*, 298(13) :305–308, 1984.
- J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l_1 -ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning, ICML '08*, pages 272–279, New York, NY, USA, 2008. Association for Computing Machinery.
- S. Dudoit and M. J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.*, 2(2) :131–154, 2005.
- G. Ecoto and A. Chambaz. Forecasting the cost of drought events in France by Super Learning. Technical report, submitted, Dec. 2022a. URL <https://hal.science/hal-03701743>.
- G. Ecoto and A. Chambaz. Forecasting the cost of drought events in france by super learning. Technical report, 2022b. URL <https://arxiv.org/abs/2206.11545>. Submitted.
- G. Ecoto, A. F. Bibaut, and A. Chambaz. One-step ahead sequential Super Learning from short times series of many slightly dependent data, and anticipating the cost of natural disasters. Technical report, submitted, July 2021a. URL <https://hal.science/hal-03300559>.
- G. Ecoto, A. F. Bibaut, and A. Chambaz. One-step ahead sequential Super Learning from short times series of many slightly dependent data, and anticipating the cost of natural disasters. Technical report, 2021b. URL <https://arxiv.org/abs/2107.13291>. Submitted.
- J. Feydy, T. Séjourné, F.-X. Vialard, S. Amari, A. Trounev, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 2019a.

J. Feydy, T. Séjourné, F.-X. Vialard, S. Amari, A. Trounev, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 16–18 Apr 2019b.

France Assureurs. Le risque sécheresse et son impact sur les habitations. 2022. URL <https://www.franceassureurs.fr/wp-content/uploads/le-risque-secheresse-et-son-impact-sur-les-habitations-15-novembre-2022-web.pdf>.

Y. Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2) :256–285, 1995.

J. H. Friedman. Greedy function approximation : A gradient boosting machine. *The Annals of Statistics*, 29(5) :1189 – 1232, 2001. doi : 10.1214/aos/1013203451. URL <https://doi.org/10.1214/aos/1013203451>.

P. Gaillard, G. Stoltz, and T. van Erven. A second-order bound with excess losses. In *Proceedings of COLT'14*, volume 35, pages 176–196. JMLR : Workshop and Conference Proceedings, 2014.

P. Gaillard, Y. Goude, L. Plagne, T. Dubois, and B. Thieurmél. *opera : Online Prediction by Expert Aggregation*, 2023. URL <http://pierre.gaillard.me/opera.html>. R package version 1.2.1.

A. Heranval, O. Lopez, and M. Thomas. Application of machine learning methods to predict drought cost in france. *European Actuarial Journal*, pages 1–23, 2022.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging : a tutorial. *Statist. Sci.*, 14(4) :382–417, 1999. With comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors.

M. Hollander and D. A. Wolfe. *Nonparametric statistical methods*. Wiley Series in Probability and Statistics : Texts and References Section. John Wiley & Sons, Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

A. E. Hubbard, S. Kherad-Pajouh, and M. J. van der Laan. Statistical inference for data adaptive target parameters. *Int. J. Biostat.*, 12(1) :3–19, 2016.

A. Iglesias, D. Assimacopoulos, and H. A. J. van Lanen, editors. *Drought : Science And Policy*. Wiley-Blackwell, aug 2019. doi : 10.1002/9781119017073.

IGN. GEOFLA. Technical report, Institut National de l’Information Géographique et Forestière, 2018. URL https://geoservices.ign.fr/sites/default/files/2021-07/DC_GEOFLA_2-2.pdf. version 2.2.

- IGN. BD TOPO. Technical report, Institut National de l'Information Géographique et Forestière, 2021. URL https://geoservices.ign.fr/sites/default/files/2021-07/DC_BDTopo_3-0.pdf. version 3.0.
- Insee. Recensement de la population 1999 : tableaux analyses. Technical report, Institut national de la statistique et des études économiques, 2000.
- S. Janson. Large deviations for sums of partly dependent random variables. *Random Structures Algorithms*, 24(3) :234–248, 2004.
- I. T. Jolliffe and D. B. Stephenson. *Forecast verification : a practitioner's guide in atmospheric science*. 2012.
- V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6) :2593–2656, 2006.
- L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband : A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185) :1–52, 2018.
- N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Inform. and Comput.*, 108(2) :212–261, 1994.
- T. Liu, A. W. Moore, and A. Gray. New algorithms for efficient high-dimensional non-parametric classification. *Journal of Machine Learning Research*, 7(41) :1135–1158, 2006.
- I. Logar and J. C. J. M. van den Bergh. Methods for assessment of the costs of droughts. Technical report, Institute of environmental science and technology, Universitat Autònoma de Barcelona, 2011. WP5 final report.
- S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- MI. Procédure de reconnaissance de l'état de catastrophe naturelle - révision des critères permettant de caractériser l'intensité des épisodes de sécheresses-réhydrations des sols à l'origine des mouvements de terrains différentiels. Technical report, Ministère de l'intérieur, 2019. URL <https://www.legifrance.gouv.fr/download/pdf/circ?id=44648>. NOR : INTE1911312C.
- MRN. Bilan annuel des principaux événements CAT-NAT & climatiques. Technical report, Mission Risques Naturels, 2023.
- MTE. Cartographie de l'exposition des maisons individuelles au retrait-gonflement des argiles. Technical report, Ministère de la Transition Écologique, 2021.

-
- A. I. Naimi and L. B. Balzer. Stacked generalization : an introduction to super learning. *Eur. J. Epidemiol.*, 33(5) :459–464, 2018. doi : 10.1007/s10654-018-0390-z.
- P. Ochs, T. Brox, and T. Pock. iPiano : inertial proximal algorithm for strongly convex optimization. *J. Math. Imaging Vision*, 53(2) :171–181, 2015.
- S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10) :1345–1359, 2010. doi : 10.1109/TKDE.2009.191.
- V. V. Petrov. *Limit theorems of probability theory*, volume 4 of *Oxford Studies in Probability*. The Clarendon Press, Oxford University Press, New York, 1995. Sequences of independent random variables, Oxford Science Publications.
- G. Peyré and M. Cuturi. Computational optimal transport, 2020.
- G. Pinto, R. Messina, H. Li, T. Hong, M. Savino Piscitelli, and A. Capozzoli. Sharing is caring : An extensive analysis of parameter-based transfer learning for the prediction of building thermal dynamics. *Energy and Buildings*, 276 :112530, 2022. ISSN 0378-7788. doi : <https://doi.org/10.1016/j.enbuild.2022.112530>. URL <https://www.sciencedirect.com/science/article/pii/S0378778822007010>.
- E. Polley, E. LeDell, C. Kennedy, and M. J. van der Laan. *SuperLearner : Super Learner Prediction*, 2021. URL <https://CRAN.R-project.org/package=SuperLearner>. R package version 2.0-28.
- E. C. Polley, S. Rose, and M. J. van der Laan. Super learning. In *Targeted learning*, Springer Ser. Statist., pages 43–66. Springer, New York, 2011.
- R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL <https://www.R-project.org/>.
- E. Rio. Moment inequalities for sums of dependent random variables under projective conditions. *J. Theoret. Probab.*, 22(1) :146–163, 2009.
- R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.
- S. Sapp, M. van der Laan, and J. Canny. Subsemble : an ensemble method for combining subset-specific algorithm fits. *Journal of Applied Statistics*, 41(6) :1247–1259, 2014. doi : 10.1080/02664763.2013.864263. URL <https://doi.org/10.1080/02664763.2013.864263>.
- A. Satriani, A. Loperte, M. Proto, and M. Bavusi. Building damage caused by tree roots : laboratory experiments of GPR and ERT surveys. *Adv. Geosci.*, 24, 2010.
- S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Mach. Learn.*, 69 :115–142, 2007.

- Swiss Re Institute. Sigma - natural catastrophes and inflation in 2022 : a perfect storm. Technical report, Swiss Re Institute, 2023.
- C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. pages 270–279, 2018.
- T. Therneau and B. Atkinson. *rpart : Recursive Partitioning and Regression Trees*, 2019. URL <https://CRAN.R-project.org/package=rpart>. R package version 4.1-15.
- M. J. van der Laan. Statistical inference for variable importance. *Int. J. Biostat.*, 2 : Art. 2, 33, 2006.
- M. J. van der Laan, S. Dudoit, and A. W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics Decisions*, 24(3) :373–395, 2006. doi : doi:10.1524/std.2006.24.3.373.
- M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6 :Art. 25, 23, 2007.
- M. Wüest, D. Bresch, and T. Corti. The hidden risks of climate change : An increase in property damage from soil subsidence in europe. Technical report, Swiss Reinsurance company Ltd, 2011. URL https://www.preventionweb.net/files/20623_soilsubsidencepublicationfinalen1.pdf.
- A. J. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *J. Amer. Math. Soc.*, 9 :1051–1094, 1996.
- B. D. Williamson, P. B. Gilbert, M. Carone, and N. Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1) :9–22, 2021.
- O. Wintenberger. Optimal learning with Bernstein online aggregation. *Mach. Learn.*, 106(1) :119–141, 2017.
- D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2) :241–259, 1992a. ISSN 0893-6080. doi : [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1). URL <https://www.sciencedirect.com/science/article/pii/S0893608005800231>.
- D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2) :241–259, 1992b.
- M. N. Wright and A. Ziegler. ranger : A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1) :1–17, 2017. doi : 10.18637/jss.v077.i01.
- J. Wu and D. Benkeser. A huber loss-based super learner with applications to healthcare expenditures. 2022.

- B. Zadrozny and C. P. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *International Conference on Machine Learning*, 2001. URL <https://api.semanticscholar.org/CorpusID:9594071>.
- J. Zhang and Y. Yang. Probabilistic score estimation with piecewise logistic regression. *Proceedings of the twenty-first international conference on Machine learning*, 2004. URL <https://api.semanticscholar.org/CorpusID:14181364>.
- F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1) :43–76, 2021. doi : 10.1109/JPROC.2020.3004555.
- M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 928–935. AAAI Press, 2003.
- M. M. E. Zumrawi, A. O. Abdelmarouf, and A. E. A. Gameil. Damages of buildings on expansive soils : Diagnosis and avoidance. *International Journal of Multidisciplinary and Scientific Emerging Research*, 6(2), May 2017.

