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

Thi Thanh Yen Nguyen, Geoffrey Ecoto, Antoine Chambaz

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

Thi Thanh Yen Nguyen[1], Geoffrey Ecoto[1,2], Antoine Chambaz[1]

[1] MAP5 (UMR CNRS 8145), Université Paris Cité
[2] Caisse Centrale de Réassurance

June 26, 2024

Abstract

Drought events rank as the second most costly natural disasters within the French legal framework of the natural disaster compensation scheme. A critical aspect of the national compensation scheme involves cities submitting requests for the government declaration of natural disaster for a drought event as a key step. We take on the challenge of forecasting which cities will submit such requests.

The problem can be tackled as a classification task, leveraging the power of classification algorithms. Taking a slightly different perspective, we introduce an alternative procedure that hinges on optimal transport theory and iPiano, 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 predictions derived from both perspectives, resulting in enhanced forecasting accuracy.

A simulation study illustrates the procedures. The real data application is presented and discussed in details. The convergence of the iPiano algorithm is established, using the notion of o-minimal structures from the field of tame geometry.

Keywords. Kurdyka-Lojasiewicz inequality, natural disaster, o-minimal structures, optimal transport, proximal algorithm, Sinkhorn algorithm, Sinkhorn divergence

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 all private property insurance policies (MTES, 2016). As 90% of the French natural disasters insurance market is reinsured by Caisse Centrale de Réassurance (henceforth abbreviated as CCR) (CCR, 2022), a public-sector reinsurer providing coverage against natural catastrophes and uninsurable risks, the French state ultimately bears 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., 2021; Ecoto and Chambaz, 2022; Ecoto et al., 2024) 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., 2021; Ecoto and Chambaz, 2022; Ecoto et al., 2024) are the only six 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 predictions derived from both perspectives.

The rest of the study is organized as follows. Section 2 introduces the data set that we obtained by merging several data sets, some of which either provided by CCR's cedents[1] 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 3 is a modicum of optimal transport theory. Section 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

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

theoretical and computational perspectives. Section 5 presents a simulation study and introduces the hybrid procedure. Section 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 7 discusses our results and outlines potential avenues for future research. In the appendix, Section A 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 o-minimal structures (Wilkie, 1996) play a central role.

2 Data and statistical challenge

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, 2022, 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 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.

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 drought events. The description of a city's exposure to drought events builds upon the SWI in a manner presented almost comprehensively in (Ecoto and Chambaz, 2022, 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 km^2 squares used by Météo-France to partition the French territory.

Note that for any year t and week $u \in \mathcal{U}_t \cap [\![44, 52]\!]$ (that is, before the end of year t), we necessarily have access to fewer than 37 values of the SWI for year t. We use a prediction model 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 September 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.

2.3 The statistical challenge and some facts about the data

As elaborated in Section 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 $\operatorname{card}\mathcal{A}_1 = \operatorname{card}\mathcal{A}_2 = 34,841$ and $\operatorname{card}\mathcal{A}_3 = 34,836$. 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 $u = \min\mathcal{U}_t$),

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 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 $\sum_{\alpha\in\mathcal{A}_t}\zeta_{\alpha,t}/\operatorname{card}\mathcal{A}_t$ $(t=1,2,3)$ 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.

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 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 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 $\sum_{\alpha\in\mathcal{A}_t}\zeta_{\alpha,t,\min\mathcal{U}_t}/\operatorname{card}\mathcal{A}_t$), 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 $\sum_{\alpha\in\mathcal{A}_t}\zeta_{\alpha,t,\max\mathcal{U}_t}/\operatorname{card}\mathcal{A}_t$), 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 $\sum_{\alpha\in\mathcal{A}_t}\zeta_{\alpha,t}/\operatorname{card}\mathcal{A}_t$), 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

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

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

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

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 (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 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%. Finally, promoting 0-predictions as part of the control of the sparsity of a set of predictions $\{\widehat{\zeta}^u_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ for a week $u \in \mathcal{U}_3$ holds merit in itself. Indeed, denoting by $\mathrm{IS}_{\alpha,3}$ the 2021 (estimated) insured sum corresponding to the houses located within the limits of any city $\alpha \in \mathcal{A}_3$ (one of the entries of $\xi_{\alpha,3,u}$, see Section 2.2), the sum

$$\sum_{\alpha \in \mathcal{A}_3} \mathrm{IS}_{\alpha,3} \, \mathbf{1}\{\zeta_{\alpha,3,u} = 1\} + \sum_{\alpha \in \mathcal{A}_3} \widehat{\zeta}^u_{\alpha,3} \mathrm{IS}_{\alpha,3} \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\} \tag{2}$$

may be used as an estimator of financial exposure due to the 2021 drought events. The contribution to (2) of a single city $\alpha \in \mathcal{A}_3$ with a large $\mathrm{IS}_{\alpha,3}$ may be significant even if its prediction $\widehat{\zeta}^u_{\alpha,3}$ is small but not 0. In addition, the contribution to (2) of many cities with moderate insured sums may be significant even if their prediction are small but not 0.

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 $\mathbf{z} := (z_1, \ldots, z_R)$ and $\mathbf{z}' := (z'_1, \ldots, z'_{R'})$ be two collections of elements of a space $\mathcal{Z}$. Let $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ 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 $[\![R]\!] \times [\![R']\!]$ with uniform marginal laws, where $[\![d]\!] := \{1, \ldots, d\}$ for every integer $d \geq 1$. For each $P \in \Pi_{R,R'}$, let

$$E(P) := - \sum_{r \in [\![R]\!], r' \in [\![R']\!]} P_{r,r'} \log P_{r,r'}$$

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

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

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

$$\mathcal{W}_\gamma(C) := \min_{P \in \Pi_{R,R'}} \left[\langle P, C \rangle - \gamma E(P) \right]. \tag{3}$$

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 (3) 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

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

(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 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.

4.1 Translation to an optimization problem

As elaborated in Section 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 (1), and so repeatedly for each $u \in \mathcal{U}_3$. In the rest of the study, it will be convenient to denote generically $\{(x_m, y_m) : m \in [\![M]\!]\} \subset \mathcal{X} \times \{0, 1\}$ and $\{(x'_n, y'_n) : n \in [\![N]\!]\} \subset \mathcal{X} \times \{0, 1\}$ two collections of couples for which it is desired to predict y'_n based on x'_n, for every $n \in [\![N]\!]$, using past observations (x_1, y_1), …, (x_M, y_M). To do so, we propose to solve the following optimization problem:

$$\operatorname*{arg\,min}_{\theta \in \mathbb{R}^N} \left\{ \mathcal{S}_\gamma(\mathbf{z}, \mathbf{z}'(\theta)) + g_\tau(\theta) \right\}, \tag{4}$$

where

- for all $\theta \in \mathbb{R}^N$,

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

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

$$c\big((x, y), (x', \theta)\big) := \mathrm{dis}(x, x')^2 + (y - \theta)^2 \tag{5}$$

  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) 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 $\mathrm{card}\{n \in [\![N]\!] : \theta_n \neq 0\}$. 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 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.

4.2 On solving (4)

Solving (4) 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 A.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.

An instance of Forward-Backward Splitting (FBS) algorithm (Attouch et al., 2010), 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)

$$\theta^{k+1} = \mathrm{Prox}_{\alpha g_\tau}\Big(\theta^k - \alpha \nabla f(\theta^k) + \beta(\theta^k - \theta^{k-1})\Big), \tag{6}$$

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

$$\mathrm{Prox}_{\alpha g_\tau}(t) := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^N} \big\{ \tfrac{1}{2}\|\theta - t\|_2^2 + \alpha g_\tau(\theta) \big\}. \tag{7}$$

On the one hand, if $g_\tau(\theta) = \tau\|\theta\|_1 + \mathbf{I}\{\theta \in [0,1]^N\}$ then (7) is simply given by

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

In particular, if $t \in [0,1]^N$ then $\left(\mathrm{Prox}_{\alpha g_\tau}(t)\right)_n = \left(t_n - \alpha\tau\right)_+$ for every $n \in [\![N]\!]$. 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 A.1.2 that the gradient of f is given by

$$\begin{aligned}\nabla f(\theta) &= \nabla\,\mathcal{W}_\gamma(C(\mathbf{z}, \mathbf{z}'(\theta)) - \tfrac{1}{2}\nabla\,\mathcal{W}_\gamma\left(C(\mathbf{z}'(\theta), \mathbf{z}'(\theta))\right) \\ &= 2(\tfrac{1}{N}\theta - \widehat{P}_\theta^\top y) - (\tfrac{2}{N}\theta - (\widehat{Q}_\theta + \widehat{Q}_\theta^\top)\theta) \\ &= -2\widehat{P}_\theta^\top y + (\widehat{Q}_\theta + \widehat{Q}_\theta^\top)\theta \end{aligned} \tag{8}$$

with

$$\widehat{P}_\theta = \operatorname*{arg\,min}_{P \in \Pi_{M,N}} \left\{\langle P, C(\mathbf{z}, \mathbf{z}'(\theta))\rangle - \gamma E(P)\right\}, \tag{9}$$

$$\widehat{Q}_\theta = \operatorname*{arg\,min}_{P \in \Pi_{N,N}} \left\{\langle P, C(\mathbf{z}'(\theta), \mathbf{z}'(\theta))\rangle - \gamma E(P)\right\}. \tag{10}$$

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 $\mathrm{dom}\, g_\tau$ and that $(f + g_\tau)$ satisfies the Kurdyka-Lojasiewicz property on its domain (the proof is presented in Section A). 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 $r(\theta) := \theta - \mathrm{Prox}_{\alpha g_\tau}(\theta - \alpha\nabla f(\theta))$, 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 A.2.1), $r(\theta) = 0$ iff

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

4.3 Implementation of the "OT-procedure"

Algorithm 1 solves (4) 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). For any vector $\theta \in \mathbb{R}^N$ and subset $\mathcal{N}$ of $[\![N]\!]$, we denote $\theta|_{\mathcal{N}} := (\theta_n)_{n\in\mathcal{N}} \in \mathbb{R}^{\mathrm{card}\,\mathcal{N}}$.

Input: Data $\{(x_m, y_m) : m \in [\![M]\!]\}$, $\{x'_n : n \in [\![N]\!]\}$; 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 $\mathcal{M} \subset [\![M]\!]$, $\mathcal{N} \subset [\![N]\!]$ of cardinality B
  Set $\mathbf{z} \leftarrow ((x_m, y_m) : m \in \mathcal{M})$ and $\mathbf{z}'(\theta^t|_{\mathcal{N}}) \leftarrow ((x'_n, \theta^t_n) : n \in \mathcal{N})$
  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 $\theta^{t+1} \leftarrow \mathrm{Prox}_{\alpha g_\tau}(\theta^{t+1})$
  Update $t \leftarrow t + 1$
end while

We wrote a `python/pytorch` program that implements Algorithm 1. Available at `https://github.com/yen-nguyen-thi-thanh/OT_prediction/tree/main`, 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 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 (5), 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 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.

5 A simple simulation study, introducing the "hybrid procedure"

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 $\mathrm{expit}(\mathrm{cst}(p) + R)$, then the joint law of (X, Y) is P_p;

- the above constant $\mathrm{cst}(p) \in \mathbb{R}$ is defined in such a way that $E_{P_p}(Y) = P_p(Y = 1) = p$.

For instance, $\mathrm{cst}(15\%) \approx -3.13$, $\mathrm{cst}(10\%) \approx -3.83$ and $\mathrm{cst}(5\%) \approx -5.00$. 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 $\ell \in [\![L]\!]$, 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 $\ell \in [\![L]\!]$, the vector $(y'_{n,\ell})_{n\in[\![N]\!]}$ based on $\{(x_{m,\ell}, y_{m,\ell}) : m \in [\![M]\!]\}$ and on $(x'_{n,\ell})_{n\in[\![N]\!]}$.

5.2 Fine-tuning the OT-procedure

Let us first describe how we fine-tune the OT-procedure in order to predict $(y'_{n,\ell})_{n\in[\![N]\!]}$ by solving (4) with $(x_m, y_m) = (x_{m,\ell}, y_{m,\ell})$ and $(x'_n, y'_n) = (x'_{n,\ell}, y'_{n,\ell})$ for all $m \in [\![M]\!]$ and $n \in [\![N]\!]$, for each $\ell \in [\![L]\!]$ in turn. On the one hand, we select the cost function $c : (\mathbb{R}^2 \times \{0, 1\}) \times (\mathbb{R}^2 \times \{0, 1\}) \to \mathbb{R}_+$ (5) 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 5.3. Furthermore, in view of Algorithm 1, we set $\gamma = 10^{-3}$, $\alpha = 10^{-3}$, $\beta = 10^{-4}$, $B = 128$ and $T = 2000$.

5.3 Alternative, classification-based approaches

As an alternative approach, we also consider training an algorithm using $\{(x_{m,\ell}, y_{m,\ell}) : m \in [\![M]\!]\}$ in order to learn to classify each $x'_{n,\ell}$ individually ($n \in [\![N]\!]$), for every $\ell \in [\![L]\!]$ 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 $\{x \mapsto \mathrm{expit}(t_0 + t_1 x_1 + t_2 x_2) : t \in \mathbb{R}^3\}$ 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 $\{x \mapsto \mathrm{expit}(t_0 + t_1 x_1 + t_2 x_2 + t_3 \sqrt{x_1^2 + x_2^2}) : t \in \mathbb{R}^4\}$ 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).

5.4 Results, introducing the "hybrid procedure"

For each $\ell \in [\![L]\!]$, we train the two super learners and denote by $\widehat{y}'^{\mathrm{SL}_1}_{n,\ell}$ and $\widehat{y}'^{\mathrm{SL}_2}_{n,\ell}$ the estimates of the conditional probabilities that $Y = 1$ given $X = x'_{n,\ell}$ that they output for each $n \in [\![N]\!]$. Next, we set $\tau = \|\widehat{y}'^{\mathrm{SL}_2}_{n,\ell}\|_1$ for the OT-procedure, run it, and denote by $\widehat{y}'^{\mathrm{OT}}_{n,\ell}$ the estimates of the conditional probability that $Y = 1$ given $X = x'_{n,\ell}$ for each $n \in [\![N]\!]$ 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 $\widehat{y}'^{\mathrm{HYB}}_{n,\ell} := (\widehat{y}'^{\mathrm{SL}_2}_{n,\ell} \times \widehat{y}'^{\mathrm{OT}}_{n,\ell})^{1/2}$ for every $n \in [\![N]\!]$.

Figure 1 provides insights into the predictions $\{\widehat{y}'^{\bullet}_{n,\ell} : n \in [\![N]\!]\}$ where the symbol $\bullet$ stands for $\mathrm{SL}_1, \mathrm{SL}_2, \mathrm{OT}, \mathrm{HYB}$. On the one hand, the empirical cumulative distribution functions (ecdfs) plotted in the left-hand side panel of Figure 1 reveal that the predictions $\widehat{y}'^{\mathrm{OT}}_{n,\ell}$ for $(n, \ell) \in [\![N]\!] \times [\![L]\!]$ 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 $\widehat{y}'^{\mathrm{SL}_1}_{n,\ell}$ and $\widehat{y}'^{\mathrm{SL}_2}_{n,\ell}$ 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 1 reveal that the predictions $\widehat{y}'^{\mathrm{OT}}_{n,\ell}$ for $(n, \ell) \in [\![N]\!] \times [\![L]\!]$ 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 $\{\widehat{y}'^{\bullet}_{n,\ell} : n \in [\![N]\!], \ell \in [\![L]\!] \text{ st } y'_{n,\ell} = y\}$ 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 $\ell \in [\![L]\!]$,

$$\mathrm{MSE}^{\bullet}_{\ell} := \frac{1}{N} \sum_{n \in [\![N]\!]} (y'_{n,\ell} - \widehat{y}'^{\bullet}_{n,\ell})^2 \tag{11}$$

where we substitute $\mathrm{SL}_1, \mathrm{SL}_2, \mathrm{OT}, \mathrm{HYB}$ for the symbol $\bullet$. The average and standard deviations of these numbers are reported in Table 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 2 allows us to go beyond comparisons in average. More that two thirds of the points are situated to the left of the

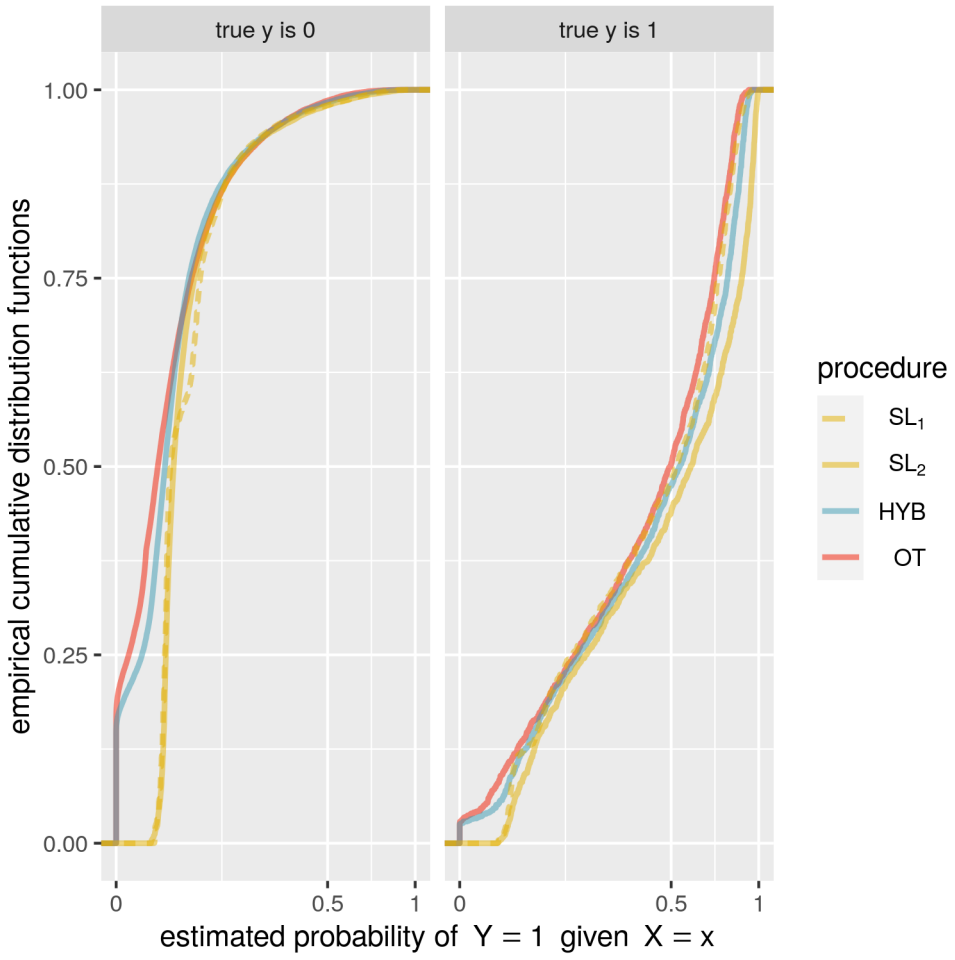

Figure 1: Empirical cumulative distribution functions of the sets $\{\widehat{y}'^{\bullet}_{n,\ell} : \ell \in [\![L]\!], n \in [\![N]\!] \text{ st } y'_{n,\ell} = y\}$ for $y = 0$ (left-hand side panel) and $y = 1$ (right-hand side panel), where the symbol $\bullet$ stands for $\mathrm{SL}_1, \mathrm{SL}_2,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

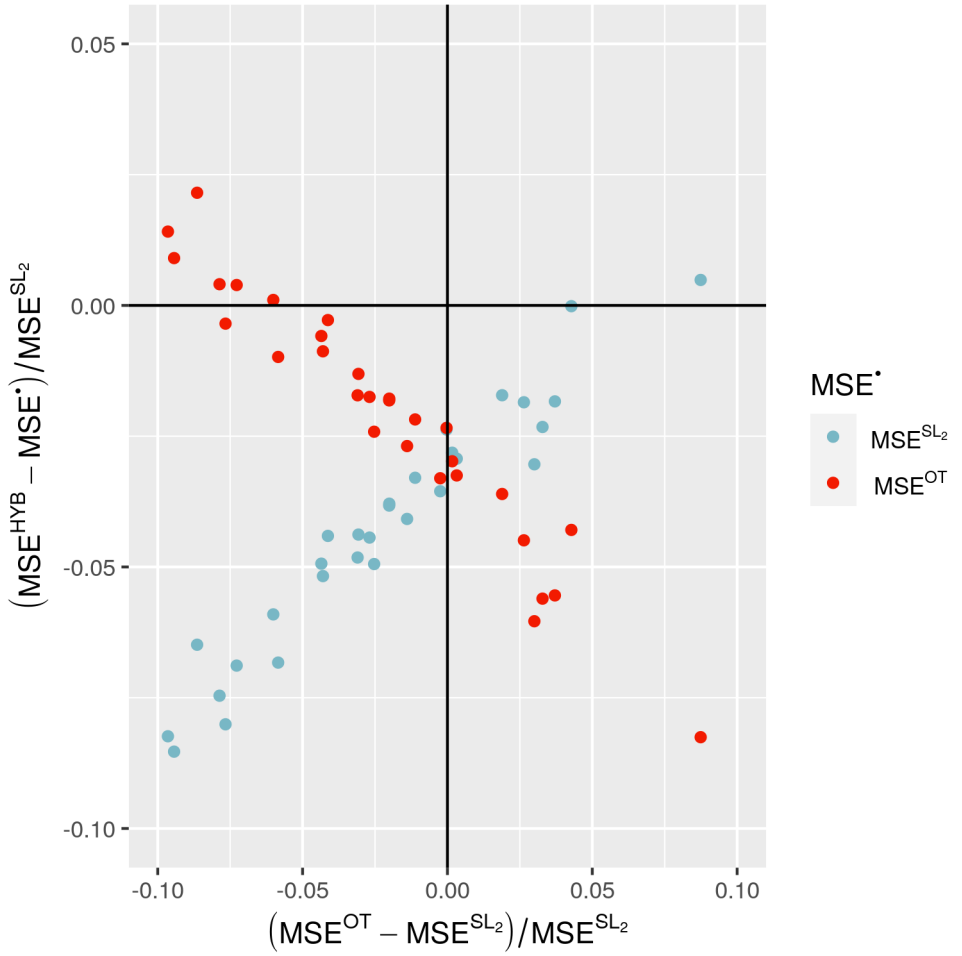

Figure 2: Scatterplot of $(\mathrm{MSE}_\ell^{\mathrm{HYB}} - \mathrm{MSE}_\ell^{\bullet})/\mathrm{MSE}_\ell^{\mathrm{SL}_2}$ against $(\mathrm{MSE}_\ell^{\mathrm{OT}} - \mathrm{MSE}_\ell^{\mathrm{SL}_2})/\mathrm{MSE}_\ell^{\mathrm{SL}_2}$ ($\ell \in [\![30]\!]$) where the symbol $\bullet$ stands for SL_2 (blue) or OT (red). See also Table 2.

black vertical line, meaning that $\mathrm{MSE}_\ell^{\mathrm{OT}}$ is smaller than $\mathrm{MSE}_\ell^{\mathrm{SL}_2}$ for the corresponding ℓs. Likewise, 29 out of 30 blue points are situated below the horizontal black line, meaning that $\mathrm{MSE}_\ell^{\mathrm{HYB}}$ is smaller than $\mathrm{MSE}_\ell^{\mathrm{SL}_2}$ for the corresponding ℓs, while 24 out of 30 red points are situated below the horizontal black line, meaning that $\mathrm{MSE}_\ell^{\mathrm{HYB}}$ is smaller than $\mathrm{MSE}_\ell^{\mathrm{OT}}$ for the corresponding ℓs. In particular, the average pattern unveiled by Table 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.

	MSE	
procedure	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 2: Averages and standard deviations of the mean squared errors $\{\mathrm{MSE}_\ell^{\bullet} : \ell \in [\![L]\!]\}$ (11) where the symbol $\bullet$ stands for $\mathrm{SL}_1, \mathrm{SL}_2, \mathrm{OT}, \mathrm{HYB}$ and $L = 30$. See also Figure 2. In each column, the smallest value stands out in bold characters.

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

6.1 Fine-tuning the OT-procedure

Defining a cost function. To begin with, we address the challenge of defining a cost function $c : (\mathcal{X} \times \mathbb{R}) \times (\mathcal{X} \times \mathbb{R}) \to \mathbb{R}_+$ (5). In view of the description of a generic vector of covariates $x \in \mathcal{X}$ made in Section 2.1, let us rewrite $x := (x_{[1]}, \ldots, x_{[4]})$ 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 $\{\xi_{\alpha,1,u} : \alpha \in \mathcal{A}_1, u \in \mathcal{U}_1\} \subset \mathcal{X}$, 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)

$$\widetilde{x} := \frac{x - \bar{\xi}_1}{\mathrm{std}_1}. \tag{12}$$

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 $x, x' \in \mathcal{X}$, $y, y' \in \mathbb{R}$,

$$c_a((x,y),(x',y')) := \sum_{k=1}^{4} a_k \|\widetilde{x}_{[k]} - \widetilde{x}'_{[k]}\|_2^2 + a_5(y - y')^2. \tag{13}$$

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), 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 (6) and Algorithm 1 in Section 4.3, we set

$$\{(x_m, y_m) : m \in [\![M]\!]\} = \{(\xi_{\alpha,1,75}, \zeta_{\alpha,1}) : \alpha \in \mathcal{A}_1 \text{ st } \zeta_{\alpha,1,75} = 0 \text{ or } (\zeta_{\alpha,1,75^-}, \zeta_{\alpha,1,75}) = (0,1)\}, \tag{14}$$

$$\{x'_n : n \in [\![N]\!]\} = \{\xi_{\alpha,2,85} : \alpha \in \mathcal{A}_2 \text{ st } \zeta_{\alpha,2,85} = 0\}, \tag{15}$$

$\alpha = 10^{-3}$, $\beta = 10^{-4}$ and $B = 128$. In words, setting (14) and (15) means that we exploit the data associated with the last week relative to year 2019 (that is, the $(75 - 52) = 23$rd 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 Successive-Halving to $\eta = 10$. For this specific couple (R, η), Hyperband consists of 4 independent "brackets"

which we present in Table 3. In the bracket indexed by $s = 0$, $n_{0,0} = 4$ different $a \in \mathbb{R}_+^5$s (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 $a \in \mathbb{R}_+^5$s are independently randomly generated; then, each is allocated $r_{s,0}$ iterations of Algorithm 1 and associated with a score. Next, recursively for $i = 1, \ldots, s$, 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 3: Resource allocations and numbers of configurations ($(r_{s,i}, n_{s,i})$, $i \in \{0, \ldots, s\}$) 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 $\widehat{\zeta}_{\alpha,2}^{\mathrm{OT},85}(a)$ 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

$$\frac{1}{N} \sum_{\alpha \in \mathcal{A}_2} (\widehat{\zeta}_{\alpha,2}^{\mathrm{OT},85}(a) - \zeta_{\alpha,2})^2 \, \mathbf{1}\{\zeta_{\alpha,2,85} = 0\}. \tag{16}$$

This completes the description of the HYPERBAND algorithm that we run to select a cost function of the form (13). Eventually, we select 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 (13) with this choice of a, we sample uniformly without replacement $M = B = 128$ elements $x_1, \ldots, x_m, \ldots, x_M$ from $\{\xi_{\alpha,1,44} : \alpha \in \mathcal{A}\} \subset \mathcal{X}$ and, independently, $N = B = 128$ elements $x'_1, \ldots, x'_n, \ldots, x'_N$ 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 (12), each x_m yields $\widetilde{x}_{m,[1]}, \widetilde{x}_{m,[2]}, \widetilde{x}_{m,[3]}, \widetilde{x}_{m,[4]}$ and each x'_n yields $\widetilde{x}'_{n,[1]}, \widetilde{x}'_{n,[2]}, \widetilde{x}'_{n,[3]}, \widetilde{x}'_{n,[4]}$. We then compute the quartiles of the sets $\{\|\widetilde{x}_{m,[k]} - \widetilde{x}'_{n,[k]}\|_2^2 : m \in [\![M]\!], n \in [\![N]\!]\}$ ($k = 1, 2, 3, 4$), which we report in Table 4.

Looking at Table 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 (13)) 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 3.

Figure 3 represents the cumulative distribution functions of the sets $\{\mathrm{cst}_{m,n} \times a_k \|\widetilde{x}_{m,[k]} - \widetilde{x}'_{n,[k]}\|_2^2 :$

$m, n \in [\![128]\!]\}$ $(k = 1, 2, 3, 4)$ where each $\mathrm{cst}_{m,n}$ (any $m, n \in [\![128]\!]$) is defined as

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

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 (13)) is driven by the corresponding groups of covariates. By this criterion, we recover the ordering suggested by Table 4.

covariates describing: $(\widetilde{x}_{[k]})$	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: Quartiles of the sets $\{\| \widetilde{x}_{m,[k]} - \widetilde{x}'_{n,[k]} \|_2^2 : m, n \in [\![128]\!]\}$ $(k = 1, 2, 3, 4)$ where $\widetilde{x}_1, \ldots, \widetilde{x}_{128}$ and $\widetilde{x}'_1, \ldots, \widetilde{x}'_{128}$ are derived from $x_1, \ldots, x_{128}$ and $x'_1, \ldots, x'_{128}$ 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 3.

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)), 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 $\tau = \bar{\zeta}_1$, (14), (15), $B = 128$ and collect the predicted probability $\widehat{\zeta}_{\alpha,2}^{\mathrm{OT},85}(\gamma, \alpha, \beta)$ 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 (16) with $\widehat{\zeta}_{\alpha,2}^{\mathrm{OT},85}(\gamma, \alpha, \beta)$ substituted for $\widehat{\zeta}_{\alpha,2}^{\mathrm{OT},85}(a)$. We select the triplet whose score is the smallest: $(\gamma, \alpha, \beta) = (10^{-2}, 10^{-3}, 10^{-4})$.

6.2 Alternative, classification-based approaches

As in the simulation study presented in Section 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 $(n \in [\![N]\!])$ using $\{(x_m, y_m) : m \in [\![M]\!]\}$. 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

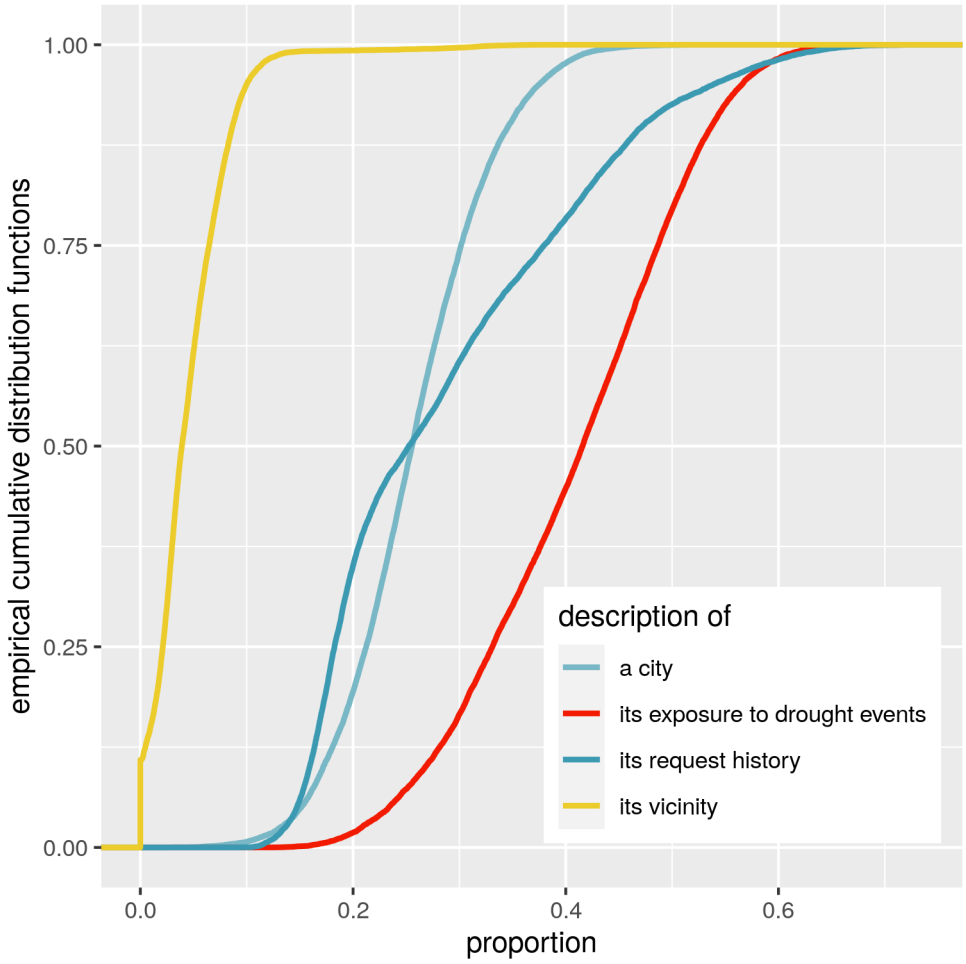

Figure 3: Cumulative distribution functions of the sets $\{\mathrm{cst}_{m,n} \times a_k \|\widetilde{x}_{m,[k]} - \widetilde{x}'_{n,[k]}\|_2^2 : m, n \in [\![128]\!]\}$ $(k = 1, 2, 3, 4)$ where $\widetilde{x}_1, \ldots, \widetilde{x}_{128}$ and $\widetilde{x}'_1, \ldots, \widetilde{x}'_{128}$ are derived from $x_1, \ldots, x_{128}$ and $x'_1, \ldots, x'_{128}$ 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 $\mathrm{cst}_{m,n}$ is such that $\mathrm{cst}_{m,n} \times \sum_{k=1}^4 a_k \|\widetilde{x}_{m,[k]} - \widetilde{x}'_{n,[k]}\|_2^2 = 1$ for all $m, n \in [\![128]\!]$. 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 (13)) is driven by the corresponding groups of covariates. See also Table 4.

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

$$\begin{aligned}&\{(x_m, y_m) : m \in [\![M]\!]\}\\ &= \{(\xi_{\alpha,1,u}, \zeta_{\alpha,1}) : \alpha \in \mathcal{A}_1, u \in \mathcal{U}_1 \text{ st } \zeta_{\alpha,1,u} = 0 \text{ or } (\zeta_{\alpha,1,u^-}, \zeta_{\alpha,1,u}) = (0,1)\},\end{aligned}$$

yielding four functions $\widehat{\zeta}_1^\bullet : \mathcal{X} \to [0,1]$, 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 $\widehat{\zeta}^{\bullet,u}_{\alpha,2} := \widehat{\zeta}^{\bullet}_{1}(\xi_{\alpha,2,u})$ 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

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

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 (we comment on the word "discrete" in the next paragraph). Lastly we retrain GLM, leveraging all data relative to years 2019 and 2020, that is

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

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

Returning to the word "discrete" mentioned in the previous paragraph, it suggests that our focus lies in determining the top-performing algorithm rather than seeking the best combination of all the algorithms. This approach is justified due to our limited hindsight, relying solely on two years of data. To illustrate, consider a future scenario where we aim to forecast the requests of the government declaration of natural disaster for a drought event for year t beyond 2021 based on data from years 2019 to $(t-1)$. The sequential learning procedure outlined above would naturally extend, opening the possibility that another algorithm may outperform GLM as the best-performing algorithm.

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 $\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3} := \widehat{\zeta}^{\mathrm{SL}}_{1:2}(\xi_{\alpha,3,u})$ 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 (13) 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 $g_\tau : \theta \mapsto \mathbf{I}\{\|\theta\|_1 \le \tau\} + \mathbf{I}\{\theta \in [0, 1]^N\}$ with $\tau = \|(\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3})_{\alpha\in\mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u}=0}\|_1$. This yields the predictions $\widehat{\zeta}^{\mathrm{OT},u}_{\alpha,3}$ 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, $\widehat{\zeta}^{\mathrm{HYB},u}_{\alpha,3} := (\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3} \times \widehat{\zeta}^{\mathrm{OT},u}_{\alpha,3})^{1/2}$ 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

$$\|(\widehat{\zeta}^{\mathrm{HYB},u}_{\alpha,3})_{\alpha\in\mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u}=0}\|_1 \le \|(\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3})_{\alpha\in\mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u}=0}\|_1 \tag{17}$$

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

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

Figure 4 shows on maps of France the probabilities $\{\widehat{\zeta}^{\mathrm{HYB},u}_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ 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 5 represents the ecdfs of the predicted probabilities $\{\widehat{\zeta}^{\bullet,u}_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ of sub-

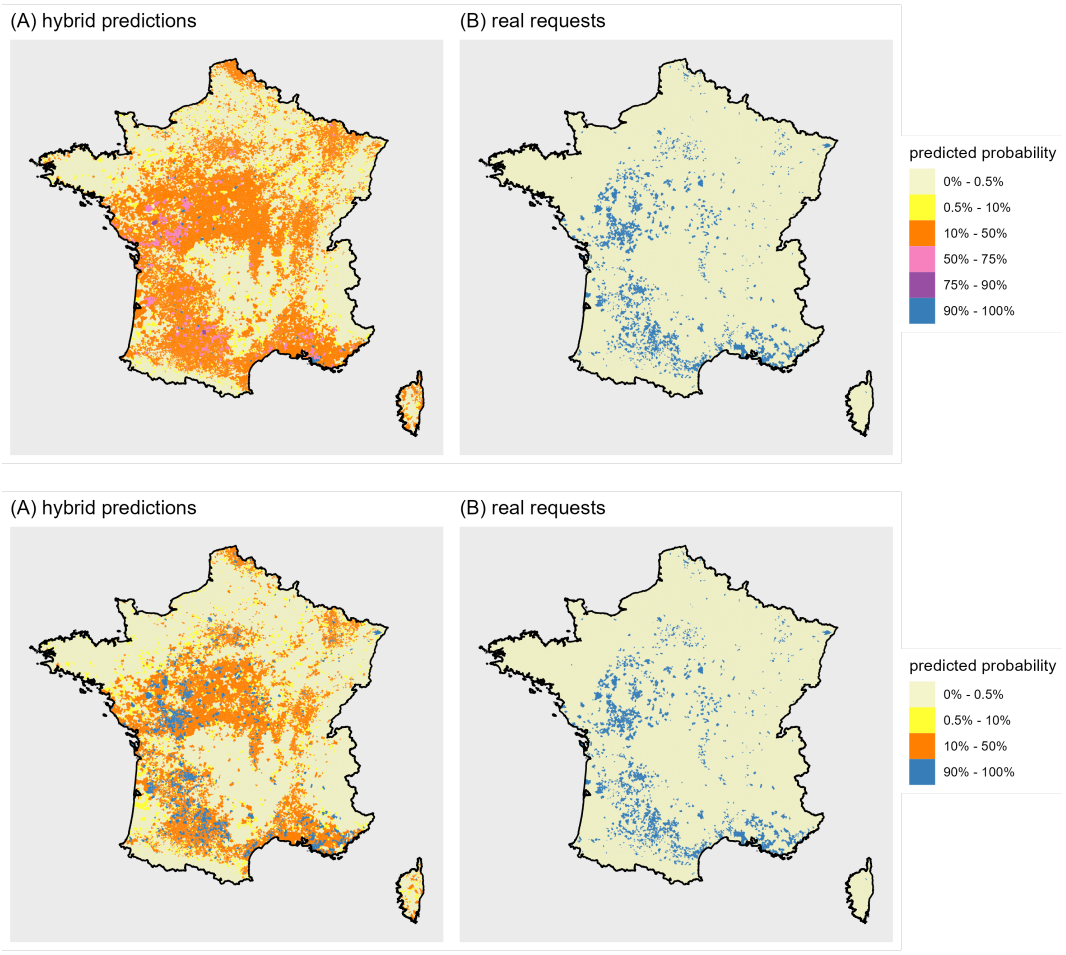

Figure 4: The left-hand side maps show the probabilities predicted by the hybrid procedure of submitting a request relative to year 2021 for weeks 49 (top) and 78 (bottom). The right-hand side maps show the cities that did submit a request eventually. In both left-hand side maps, the cities for which 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.

mitting 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 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 6 complements Figure 5 by providing a ROC (Receiver Operating Characteristic) perspective. The figure clearly demonstrates that both the OT- and hybrid procedures yield predicted probabilities predominantly below 50%. Furthermore, zooming into the left part of the graphs, specifically examining false positive rates (FPR) between 0 and 50%, and concentrating on the 49th, 59th and 68th weeks, the figure reveals a slight superiority of the hybrid procedure, in the sense that its curve consistently appears atop the others.

Figure 7 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 $\widehat{\zeta}^{\mathrm{OT},u}_{\alpha,3} > \widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3}$. 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 $\{\widehat{\zeta}^{\mathrm{OT},u}_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ are smaller than those of $\{\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$ 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 $\{\widehat{\zeta}^{\mathrm{OT},u}_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } (\zeta_{\alpha,3,u}, \zeta_{\alpha,3,u^-}) = (1,0)\}$ is larger than that of $\{\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } (\zeta_{\alpha,3,u}, \zeta_{\alpha,3,u^-}) = (1,0)\}$ 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 8 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

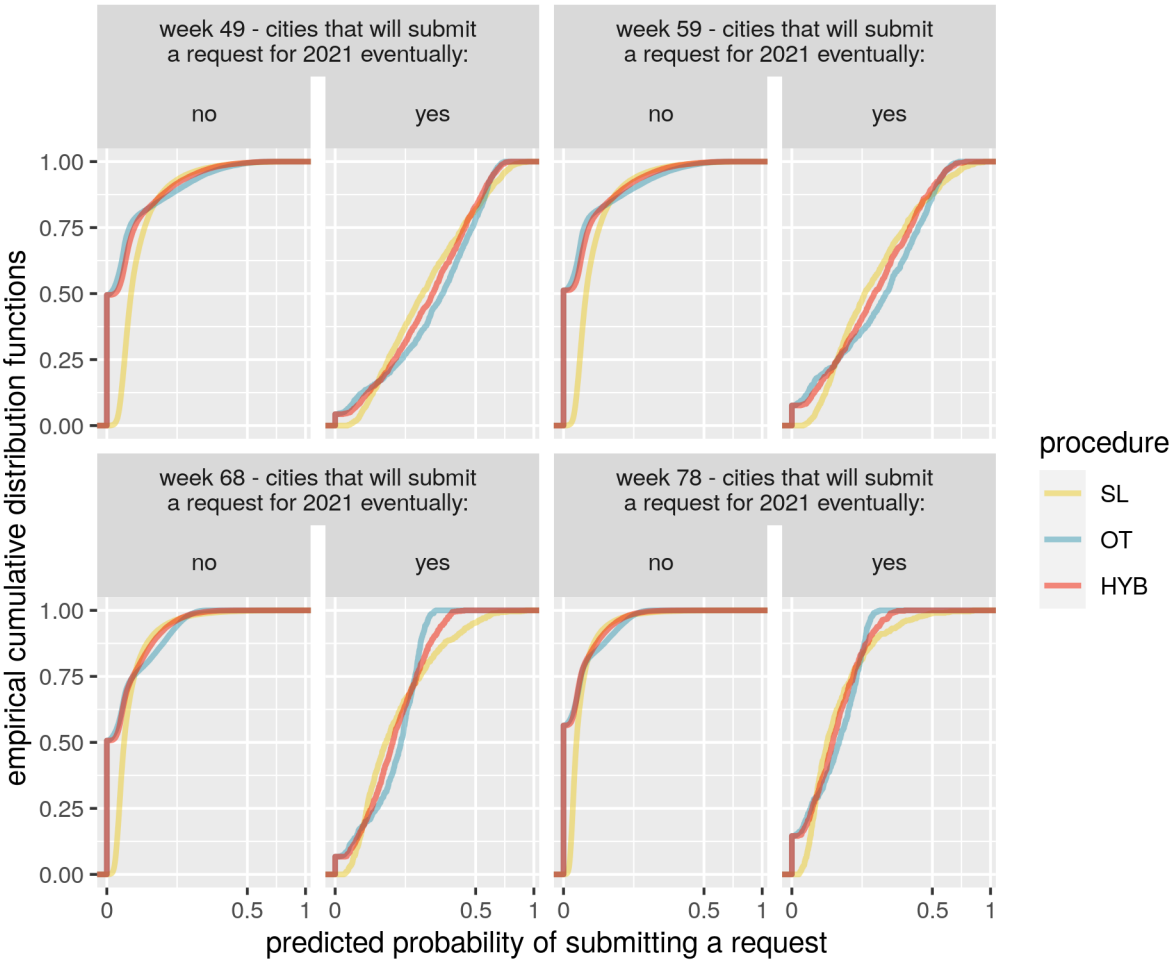

Figure 5: This plot shows, when week u is one of the 49th week of 2021 (December 6th to 12th), the $(59-52)=7$th, $(69-52)=17$th and $(78-52)=26$th weeks of 2022 (February 15th to 21st, April 26th to May 2nd, June 28th to July 4th), the empirical cumulative distribution functions (ecdfs) of the predicted 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 $\{\widehat{\zeta}^{\bullet,u}_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3}=0\}$, left-hand side panels) and for those that will (that is, the ecdfs of $\{\widehat{\zeta}^{\bullet,u}_{\alpha,3} : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3}=1\}$, right-hand side panels). See also Figure 6 for a ROC perspective and Figure 8 for a focus on medians.

than that of the predictions made by the hybrid procedure which, in turn, is consistently larger than that of the predictions made by the OT-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 the predictions made by the hybird procedure which, in turn, is consistently smaller than that of the predictions made by the OT-procedure. From this perspective, the OT-procedure outperforms the hybrid procedure which, in turn, performs better than the super learner.

To conclude, we report in Table 5 the week-specific MSE scores

$$\frac{\sum_{\alpha\in\mathcal{A}_3}(\widehat{\zeta}^{\bullet,u}_{\alpha,3}-\zeta_{\alpha,3})^2\,\mathbf{1}\{\zeta_{\alpha,3,u}=0\}}{\sum_{\alpha\in\mathcal{A}_3}\mathbf{1}\{\zeta_{\alpha,3,u}=0\}} \tag{18}$$

(all $u\in\mathcal{U}_3$, the symbol $\bullet$ standing for SL, OT and HYB). The key insight from Table 5 is that the hybrid procedure exhibits superior performance, by consistently outperforming both the OT-procedure

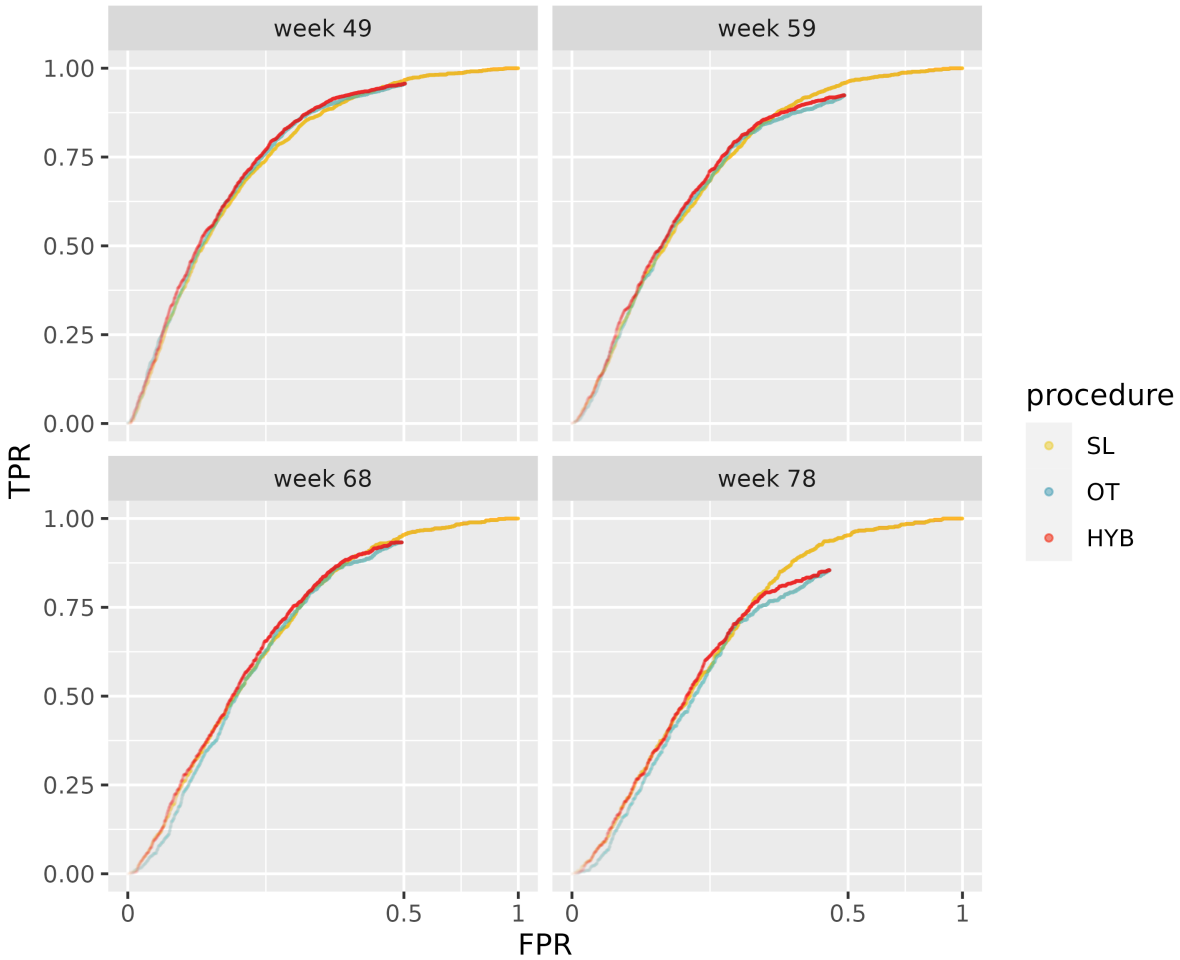

Figure 6: This plot shows, when week u is one of the 49th week of 2021 (December 6th to 12th), the $(59-52)=7$th, $(69-52)=17$th and $(78-52)=26$th weeks of 2022 (February 15th to 21st, April 26th to May 2nd, June 28th to July 4th), ROC-like curves for the predicted probabilities of submitting a request made by procedures SL, OT and HYB. FPR and TPR stand for False Positive Rate and True Positive Rate, respectively. See also Figure 5 for a focus on cities that will or that will not eventually submit a request for the government declaration of natural disaster and Figure 8 for a focus on medians.

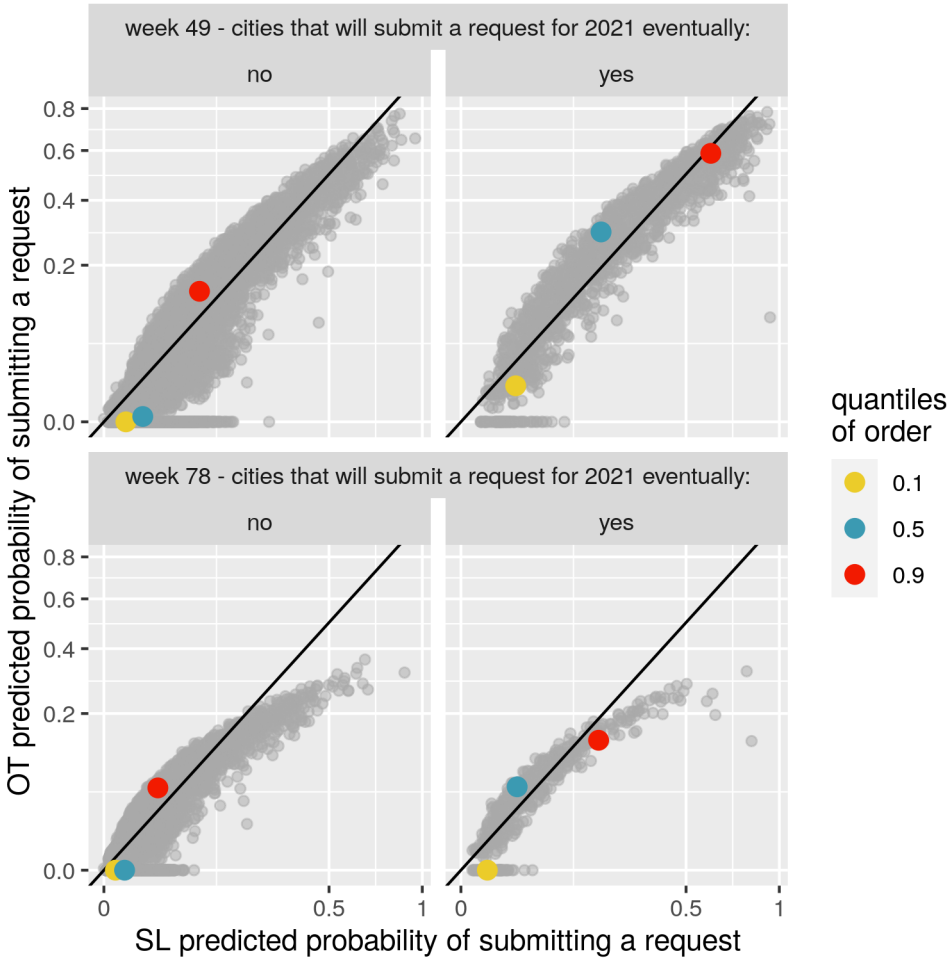

Figure 7: This plot shows, for week u equal either to the 49th week of 2021 (December 6th to December 12th) or the $(78-52) = 26$th week of 2022 (June 27th to July 3rd), the predicted probabilities of submitting a request made by procedures SL (x-axis) and OT (y-axis) separately for those cities that will not eventually submit a request for the government declaration of natural disaster for a drought event for year 2021 (that is, $\{(\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3}, \widehat{\zeta}^{\mathrm{OT},u}_{\alpha,3}) : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 0\}$, left-hand side panels) and for those that will (that is, $\{(\widehat{\zeta}^{\mathrm{SL},u}_{\alpha,3}, \widehat{\zeta}^{\mathrm{OT},u}_{\alpha,3}) : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 1\}$, right-hand side panels). In addition, three colored points represent in each panel the coordinate-specific quantiles of order 10%, 50% and 90%.

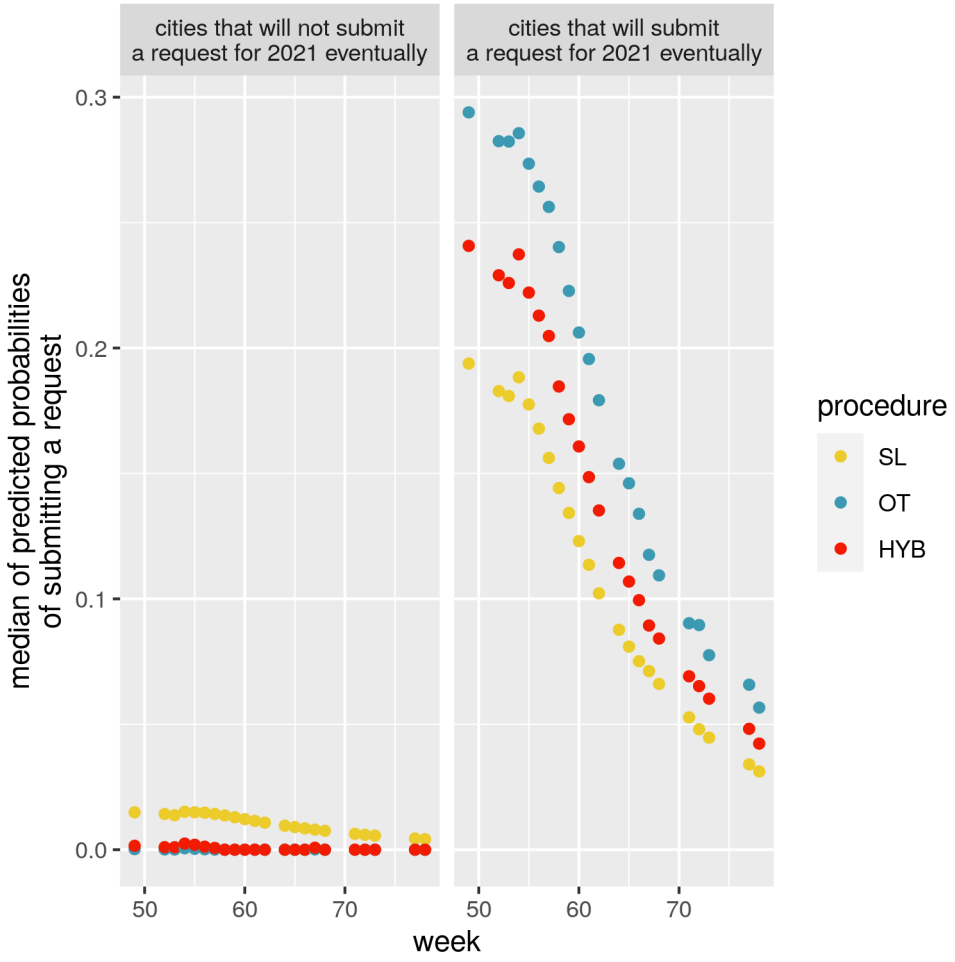

Figure 8: This plot shows, as week u goes from the 49th week of 2021 (December 6th to December 12th) to the $(78-52) = 26$th week of 2022 (June 27th to July 3rd), the evolutions of the medians of the predicted probabilities of submitting a request made by procedures SL, OT and HYB separately for those 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 $u \mapsto \text{median}\{\widehat{\zeta}^{\bullet,u}_{\alpha,3} : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 0\}$, left-hand side panel) and for those that will (that is, of $u \mapsto \text{median}\{\widehat{\zeta}^{\bullet,u}_{\alpha,3} : \alpha \in \mathcal{A}_t \text{ st } \zeta_{\alpha,3,u} = 0, \zeta_{\alpha,3} = 1\}$, right-hand side panel). See also Figure 5 for more comprehensive descriptions through empirical cumulative distribution functions.

	MSE				MSE		
week u	SL	OT	HYB	week u	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 5: Evolution of MSE $u \mapsto n_{3,u}^{-1} \sum_{\alpha \in \mathcal{A}_3} (\widehat{\zeta}_{\alpha,3}^{\bullet,u} - \zeta_{\alpha,3})^2 \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ 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 9.

and the super learner. Interestingly we also observe that, for every procedure, (18) 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 (18) for $u \in \mathcal{U}_3$ is represented in Figure 9, 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 ($u \mapsto \sum_{\alpha \in \mathcal{U}_3} \widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u} \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$). The quartiles and range of

$$\left\{ \sum_{\alpha \in \mathcal{A}_3} \zeta_{\alpha,3,u} + \sum_{\alpha \in \mathcal{U}_3} \widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u} \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\} : u \in \mathcal{U}_3 \right\} \tag{19}$$

(the heights of the bars in Figure 9) 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

$$\left\{ \sum_{\alpha \in \mathcal{A}_3} \zeta_{\alpha,3,u} + \sum_{\alpha \in \mathcal{U}_3} \widehat{\zeta}_{\alpha,3}^{\mathrm{SL},u} \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\} : u \in \mathcal{U}_3 \right\} \tag{20}$$

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 (17), it was guaranteed that each of the quartile and mean associated to (19) would be smaller then its counterpart associated to (20). Both convex hulls of (19) and (20) 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 (19) encompassing the true value, unlike that of (20).

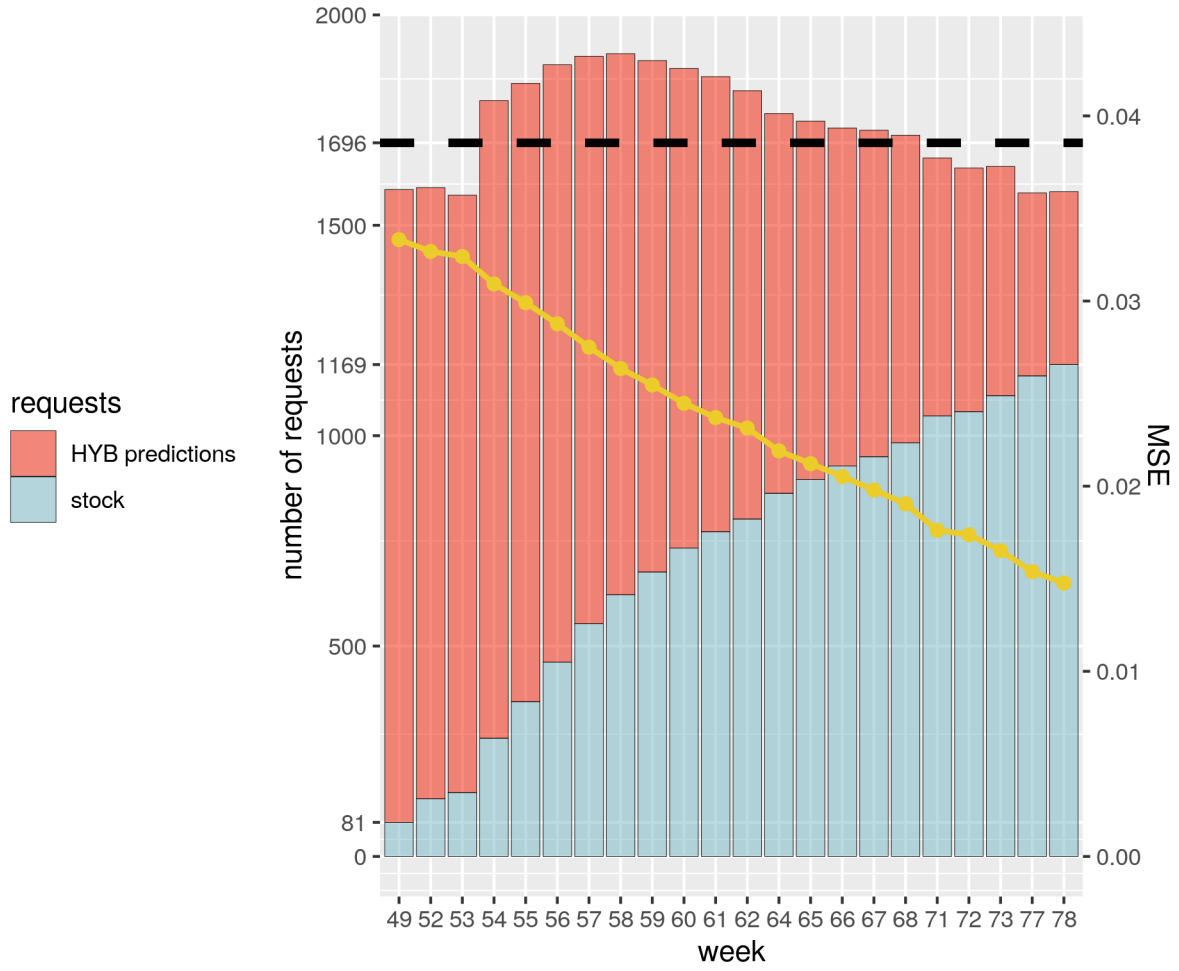

Figure 9: This plot shows, as week u goes from the 49th week of 2021 (December 6th to 12th) to the $(78 - 52) = 26$th week of 2022 (June 27th to July 3rd), the evolutions of the cardinality of the stock of requests already submitted for the government declaration of natural disaster for a drought event 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 $u \mapsto \sum_{\alpha \in \mathcal{A}_t} \widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u} \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$, 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 $u \mapsto n_{3,u}^{-1} \sum_{\alpha \in \mathcal{A}_3} (\widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u} - \zeta_{\alpha,3})^2 \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}$ 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 5.

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 sth covariate) in a generic $\xi_{\alpha,3,u}$ has on the prediction $\widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u}$ 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 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, 2022, Section 4.4).

Set arbitrarily $s \in [\![67]\!]$ 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 $\{(\widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u}, \xi_{\alpha,3,u,s}) : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$:

$$\rho_s^u := \left(\frac{\sum_{\nu=1}^{v} n_\nu (\bar{\zeta}_\nu - \bar{\zeta})^2}{\sum_{\alpha \in \mathcal{A}_3} (\widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u} - \bar{\zeta})^2 \, \mathbf{1}\{\zeta_{\alpha,3,u} = 0\}} \right)^{1/2}$$

  where n_ν is the number of $\xi_{\alpha,3,u,s}$ equal to ν, $\bar{\zeta}_\nu$ is the average of the $\widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u}$s such that $\xi_{\alpha,3,u,s} = \nu$ and $\bar{\zeta}$ is the average of all $\widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u}$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 $\{(\widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u}, \xi_{\alpha,3,u,s}) : \alpha \in \mathcal{A}_3 \text{ st } \zeta_{\alpha,3,u} = 0\}$.

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^us 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 10 shows the evolutions of $u \mapsto \rho_s^u$ for every eligible $s \in [\![67]\!]$, 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,

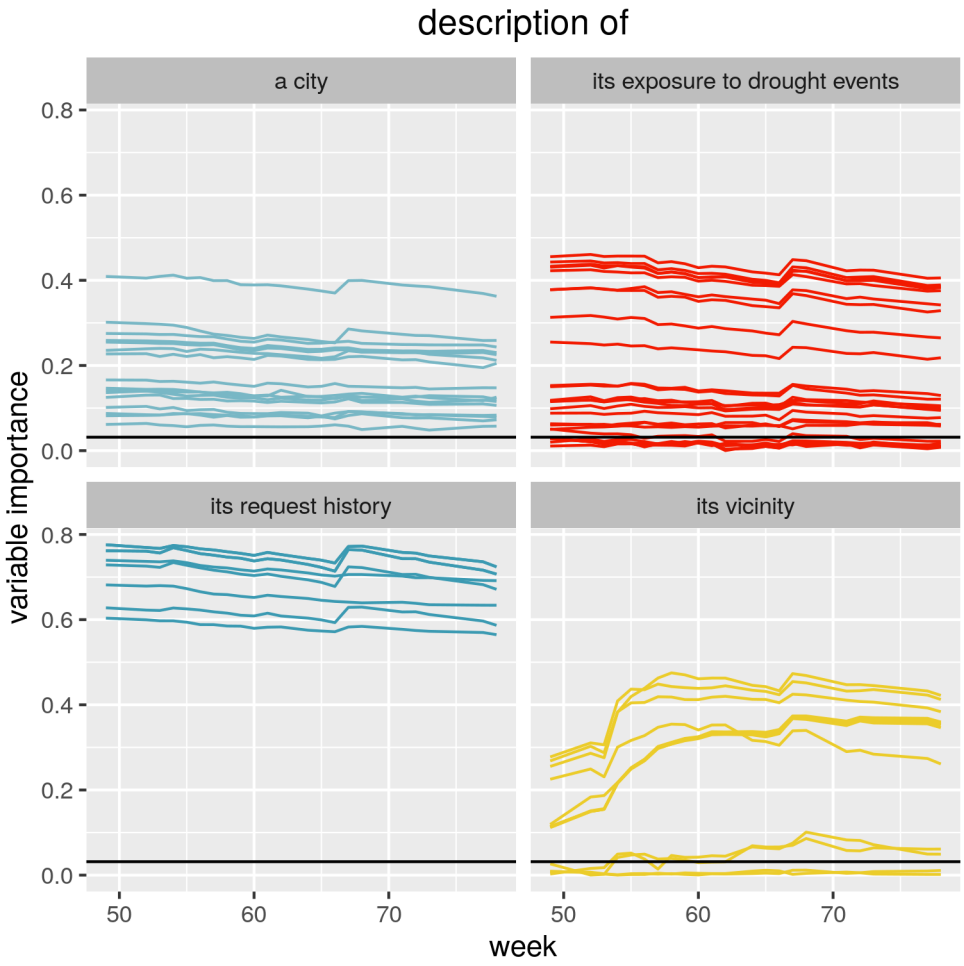

Figure 10: This plot shows, as week u goes from the 49th week of 2021 (December 6th to 12th) to the $(78-52)=26$th week of 2022 (June 27th to July 3rd), the evolutions of the importance of each variable used to make predictions, as defined in Section 6.4. For every eligible $s \in [\![67]\!]$, the larger is ρ_s^u, the stronger is the association between the sth covariate $\xi_{\alpha,3,u,s}$ and the prediction $\widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u}$ 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 6.

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 6, we report the five variables which, in each group of covariates, feature the largest average variable importance $(\sum_{u \in \mathcal{U}_3} \rho_s^u / \operatorname{card} \mathcal{U}_3)$.

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.

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 6: The five variables used to make predictions with the highest average importance ($\sum_{u\in\mathcal{U}_3}\rho_s^u/\operatorname{card}\mathcal{U}_3$, see definition in Section 6.4) in each group of covariates. For every eligible $s\in[\![67]\!]$, the larger is ρ_s^u, the stronger is the association between the sth covariate $\xi_{\alpha,3,u,s}$ and the prediction $\widehat{\zeta}_{\alpha,3}^{\mathrm{HYB},u}$ across $\alpha\in\mathcal{A}_3$ such that $\zeta_{\alpha,3,u}=0$. See also Figure 10.

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.

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 o-minimal structures from the field of tame geometry (Wilkie, 1996) to prove that a critical function related to (4) 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 `https://github.com/yen-nguyen-thi-thanh/OT_prediction/tree/main`.

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 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, we could use cross-fitting to set the value of τ prior to running Algorithm 1 (see the first paragraph of Section 6.3). Thirdly, 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. Fourthly, the current predictions obtained from the investigated procedures lack a measure of confidence. Developping 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.

Acknowledgments. The authors wish to thank Rémy Abergel (MAP5, UMR CNRS 8145, Université Paris Cité) and Jérôme Bolte (Toulouse School of Economics) for valuable discussions.

References

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.

R. Benedetti and J-J. Risler. *Real algebraic and semi-algebraic sets.* Actualités Mathématiques. [Current Mathematical Topics]. Hermann, Paris, 1990.

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.

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. Rapport d'activité 2021. Technical report, Caisse Centrale de Réassurance, 2022. URL `https://www.ccr.fr/documents/35794/35839/CCR+RA+2021+web+all+24032022.pdf/84e4c7da-34b5-22e0-e048-06a0836b7392?t=1648135815072`.

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`.

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. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l1-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.

G. Ecoto and A. Chambaz. Forecasting the cost of drought events in France by Super Learning. Technical report, submitted, December 2022. URL `https://hal.science/hal-03701743`.

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 2021. URL `https://hal.science/hal-03300559`.

G. Ecoto, A. F. Bibaut, and A. Chambaz. Forecasting the cost of drought events in france by super learning from a short time series of many slightly dependent data. *Computational Statistics*, 2024. Accepted for publication.

J. Feydy, T. Séjourné, F-X. Vialard, S. Amari, A. Trouvé, 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. Trouvé, 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`.

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.

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.

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.

Insee. Recensement de la population 1999: tableaux analyses. Technical report, Institut national de la statistique et des études économiques, 2000.

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.

T. Liu, A. W. Moore, and A. Gray. New algorithms for efficient high-dimensional nonparametric 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. Von 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 a l'origine des mouvement de terrains différentiels. Technical report, Ministère de l'intérieur, 2019. URL `https://www.legifrance.gouv.fr/download/pdf/circ?id=44648`. NOR: INTE1911312C.

MTES. Le retrait-gonflement des argiles: comment prévenir les désordres dans l'habitat individuel. Technical report, Ministère de la transition écologique et solidaire, 2016. URL `https://www.ecologie.gouv.fr/sites/default/files/dppr_secheresse_v5tbd.pdf`.

P. Ochs, T. Brox, and T. Pock. iPiano: inertial proximal algorithm for strongly convex optimization. *J. Math. Imaging Vision*, 53(2):171–181, 2015.

G. Peyré and M. Cuturi. Computational optimal transport, 2020.

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/`.

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.

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, E. C. Polley, and A. E. Hubbard. Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6:Art. 25, 23, 2007.

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.

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.

A Appendix: checking the iPiano assumptions

The iPiano assumptions consist in

1. f being C^1-smooth with a Lipschitz continuous gradient on $\operatorname{dom} g_\tau$, see Section A.1;
2. for any $\delta > 0$, $H_\delta : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ 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 $(\theta^\star, \theta^\star)$ of the sequence $(\theta^k)_{k\geq 1}$, see Section A.2.

A.1 The function f is C^1-smooth and its gradient is Lipschitz continuous on $\operatorname{dom} g_\tau$

A.1.1 Preliminaries

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 $\|A\|_F := \left(\sum_{i\in[\![d]\!],j\in[\![d']\!]} A_{i,j}^2\right)^{1/2}$ and $\|A\|_{\max} := \max\{|A_{i,j}| : i \in [\![d]\!], j \in [\![d']\!]\}$. For any vector $x \in \mathbb{R}^d$, the variation seminorm of x is defined as $\|x\|_{\mathrm{var}} := \max\{x_i : i \in [\![d]\!]\} - \min\{x_i : i \in [\![d]\!]\}$. We will use the following classical inequalities and equality:

$$\forall A \in \mathbb{R}^{d\times d'}, \forall B \in \mathbb{R}^{d'\times d''}, \|AB\|_F \leq \|A\|_F\|B\|_F; \tag{21}$$

$$\forall A \in \mathbb{R}^{d\times d'}, \forall x \in \mathbb{R}^{d'}, \|Ax\|_2 \leq \|A\|_F\|x\|_2; \tag{22}$$

$$\forall x \in \mathbb{R}^d, \|\operatorname{diag}(x)\|_F = \|x\|_2; \tag{23}$$

$$\forall x \in \mathbb{R}^d, \|x\|_{\mathrm{var}} \leq 2\|x\|_\infty; \tag{24}$$

$$\forall x \in \{0\} \times \mathbb{R}^{d-1}, \|x\|_\infty \leq \|x\|_{\mathrm{var}}. \tag{25}$$

On the Hilbert projective metric. The Hilbert projective metric on $(\mathbb{R}_+^*)^d$ is defined by

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') := \log \max\left\{\frac{x_i x'_j}{x'_i x_j} : i, j \in [\![d]\!]\right\}.$$

We will use the following properties (Birkhoff, 1957):

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') = \|\log(x) - \log(x')\|_{\mathrm{var}}; \tag{26}$$

$$\forall x, x' \in (\mathbb{R}_+^*)^d, d_{\mathcal{H}}(x, x') = d_{\mathcal{H}}(x/x', \mathbf{1}_d) = d_{\mathcal{H}}(\mathbf{1}_d/x', \mathbf{1}_d/x); \tag{27}$$

$$\forall K \in (\mathbb{R}_+^*)^{d\times d'}, \forall x, x' \in (\mathbb{R}_+^*)^{d'}, d_{\mathcal{H}}(Kx, Kx') \leq \lambda(K) d_{\mathcal{H}}(x, x'), \tag{28}$$

where $\lambda(K) := \frac{\sqrt{\eta(K)}-1}{\sqrt{\eta(K)}+1} < 1$ with $\eta(K) := \max\left\{\frac{K_{i,k}K_{j,\ell}}{K_{j,k}K_{i,\ell}} : i, j \in [\![d]\!], k, \ell \in [\![d']\!]\right\}$.

We end this section with a lemma.

Lemma 1. *Let $x, x' \in (\mathbb{R}_+^*)^d$ be such that $0 < t \leq \min\{x_j, x'_j : j \in [\![d]\!]\} \leq \max\{x_j, x'_j : j \in [\![d]\!]\} \leq T$. It holds that $\frac{1}{2} t d_{\mathcal{H}}(x, x') \leq \|x - x'\|_2$. Moreover, if $x_1 = x'_1 = 1$, then it also holds that $\|x - x'\|_2 \leq \sqrt{d} T d_{\mathcal{H}}(x, x')$.*

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 $i \in [\![d]\!]$. 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

$$\begin{aligned}
|x_i - x'_i| &= \max(x_i, x'_i) \times |1 - e^{-|\ell_i - \ell'_i|}| \\
&\leq T \times |\ell_i - \ell'_i| \quad \text{because } |1 - e^{-|q|}| \leq |q| \text{ for all } q \in \mathbb{R} \\
&\leq T \times \|\ell - \ell'\|_\infty \\
&\leq T \times \|\ell - \ell'\|_{\text{var}} \quad \text{by (25) since } \ell_1 = \ell'_1 = 0 \\
&= T d_{\mathcal{H}}(x, x') \quad \text{by (26).}
\end{aligned}$$

Consequently, $\|x - x'\|_2 \leq \sqrt{d}\|x - x'\|_\infty \leq \sqrt{d} T d_{\mathcal{H}}(x, x')$. 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

$$\begin{aligned}
\|x - x'\|_2 \geq \|x - x'\|_\infty \geq t\|\ell - \ell'\|_\infty \geq \tfrac{1}{2} t \|\ell - \ell'\|_{\text{var}} \quad &\text{by (24)} \\
= \tfrac{1}{2} t d_{\mathcal{H}}(x, x') \quad &\text{by (26).}
\end{aligned}$$

This completes the proof. □

A.1.2 The function f is differentiable

To prove that f is differentiable, we rely on the following classical result (Danskin, 1966):

Theorem 1 (Danskin's theorem, Proposition B.25 in Bertsekas (1999))**.** *Let $\mathcal{C} \subset \mathbb{R}^{d'}$ be a compact set and $\phi : \mathbb{R}^d \times \mathcal{C} \to \mathbb{R}$ be a continuous function such that $\phi(\cdot, y)$ is convex for every $y \in \mathcal{C}$. The function $\psi : \mathbb{R}^d \to \mathbb{R}$ 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 $\phi : \mathbb{R}^{R \times R'} \times \Pi_{R,R'} \to \mathbb{R}$ 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 $\psi : \mathbb{R}^{R \times R'} \to \mathbb{R}$ 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 $\widehat{P}_C$ such that $\psi(C) = \phi(C, \widehat{P}_C)$ (Cuturi and Doucet, 2014, Proposition 4.3) and $\phi(\cdot, \widehat{P}_C)$ 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 $\nabla \mathcal{W}_\gamma(C) = \widehat{P}_C$.

We use now that $f = f_a - \frac{1}{2} f_b + \text{constant}$ with $f_a, f_b : \mathbb{R}^N \to \mathbb{R}$ given by

$$f_a(\theta) := \mathcal{W}_\gamma\left(C(\mathbf{z}, \mathbf{z}'(\theta))\right) \quad \text{and} \quad f_b(\theta) := \mathcal{W}_\gamma\left(C(\mathbf{z}'(\theta), \mathbf{z}'(\theta))\right)$$

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

$$\nabla f_a(\theta) = 2(\tfrac{1}{N}\theta - \widehat{P}_\theta^\top y) \quad \text{and} \quad \nabla f_b(\theta) = 2(\tfrac{2}{N}\theta - (\widehat{Q}_\theta + \widehat{Q}_\theta^\top)\theta)$$

($\widehat{P}_\theta$ and $\widehat{Q}_\theta$ are defined in (9) and (10)). Therefore f is differentiable at every $\theta \in \mathbb{R}^N$ and (8) follows straightforwardly.

A.1.3 $\widehat{P}_\theta$ and $\widehat{Q}_\theta$ 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 $\operatorname{dom} g_\tau$ is a consequence of the following lemma.

Lemma 2. *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 (3) 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 A.1.2) are obviously bounded and Lipschitz continuous.

Let us prove Lemma 2. By (Cuturi and Doucet, 2014, Proposition 4.3), for every $\theta \in \mathbb{R}^{R'}$,

$$\widehat{P}(\theta) = \operatorname{diag}(\widehat{u}(\theta)) K(\theta) \operatorname{diag}(\widehat{v}(\theta)),$$

where $\widehat{u} : \mathbb{R}^{R'} \to (\mathbb{R}_+^*)^R$, $\widehat{v} : \mathbb{R}^{R'} \to (\mathbb{R}_+^*)^{R'}$ and the Gibbs kernel functions $K : \mathbb{R}^{R'} \to \mathbb{R}^{R\times R'}$, given by

$$K(\theta) := \Big(\exp\Big[-\left(C(\theta)\right)_{r,r'}/\gamma\Big]\Big)_{r\in[\![R]\!], r'\in[\![R']\!]}$$

satisfy the mass conservation constraints inherent to $\Pi_{R,R'}$:

$$\operatorname{diag}(\widehat{u}(\theta)) K(\theta) \operatorname{diag}(\widehat{v}(\theta))\, \mathbf{1}_{R'} = \tfrac{1}{R}\, \mathbf{1}_R \tag{29}$$

$$\operatorname{diag}(\widehat{v}(\theta)) K(\theta)^\top \operatorname{diag}(\widehat{u}(\theta))\, \mathbf{1}_{R} = \tfrac{1}{R'}\, \mathbf{1}_{R'}, \tag{30}$$

Equivalently, using the entrywise division of vectors,

$$\widehat{u}(\theta) = \frac{\frac{1}{R}\,\mathbf{1}_R}{K(\theta)\widehat{v}(\theta)}, \quad \widehat{v}(\theta) = \frac{\frac{1}{R'}\,\mathbf{1}_{R'}}{K(\theta)^\top \widehat{u}(\theta)}. \tag{31}$$

Note that $(\rho\widehat{u}(\theta), \widehat{v}(\theta)/\rho)$ also satisfy (29) and (30) for any $\rho > 0$. Thus, without loss of generality, we can impose from now on that, for all $\theta \in \operatorname{dom} g_\tau$, 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 $\operatorname{dom} g_\tau$ with Lipschitz constant $L_K := k_u^2 L_C^2/\gamma^2$ where $k_u := \max\{(K(\theta))_{r,r'} : r \in [\![R]\!], r' \in [\![R']\!]\}$ 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 $\theta, \theta' \in [0,1]^{R'}$, $r \in [\![R]\!]$ and $r' \in [\![R']\!]$, it holds that

$$\begin{aligned}
&|(K(\theta))_{r,r'} - (K(\theta'))_{r,r'}| \\
&= \max\{e^{-(C(\theta))_{r,r'}/\gamma}, e^{-(C(\theta'))_{r,r'}/\gamma}\} \times |1 - \exp(-|(C(\theta))_{r,r'} - (C(\theta'))_{r,r'}|/\gamma)|
\end{aligned}$$

$$\leq \frac{k_u}{\gamma} \times |(C(\theta))_{r,r'} - (C(\theta'))_{r,r'}|.$$

Therefore,

$$\begin{aligned}\|K(\theta) - K(\theta')\|_F^2 &= \sum_{r\in[\![R]\!],r'\in[\![R']\!]} [(K(\theta))_{r,r'} - (K(\theta'))_{r,r'}]^2 \\ &\leq \frac{k_u^2}{\gamma^2} \sum_{r\in[\![R]\!],r'\in[\![R']\!]} [(C(\theta))_{r,r'} - (C(\theta'))_{r,r'}]^2 \\ &\leq \frac{k_u^2 L_C^2}{\gamma^2}\|\theta - \theta'\|_2^2.\end{aligned}$$

- Denote $k_\ell := \min\{(K(\theta))_{r,r'} : r \in [\![R]\!], r' \in [\![R']\!]\}$. For every $\theta \in \operatorname{dom} g_\tau$,

$$\lambda(K(\theta)) \leq \Lambda := (k_u - k_\ell)/(k_u + k_\ell) < 1. \tag{32}$$

Proof: Because $k_\ell \leq (K(\theta))_{r,r'} \leq k_u$ for all $\theta \in \operatorname{dom} g_\tau$, $r \in [\![R]\!], r' \in [\![R']\!]$, 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 $i,j \in [\![R]\!], k,\ell \in [\![R']\!]$. 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 $\theta \in \operatorname{dom} g_\tau$, $\widehat{u}(\theta)$ and $\widehat{v}(\theta)$ are uniformly bounded: for all $r \in [\![R]\!]$, $r' \in [\![R']\!]$,

$$\frac{k_\ell}{k_u R'} \leq \widehat{u}_r(\theta) \leq \frac{k_u R}{k_\ell}, \tag{33}$$

$$\frac{k_\ell}{k_u^2 R' R^2} \leq \widehat{v}_{r'}(\theta) \leq \frac{1}{k_\ell R}. \tag{34}$$

Proof: Set arbitrarily $\theta \in \operatorname{dom} g_\tau$. In view of (29) (first row), since $\widehat{u}_1(\theta) = 1$, we have

$$k_\ell \|\widehat{v}(\theta)\|_\infty \leq \frac{1}{R} = \sum_{r'\in[\![R']\!]} (K(\theta))_{1r'}\widehat{v}_{r'}(\theta) \leq k_u R' \|\widehat{v}(\theta)\|_\infty. \tag{35}$$

Set $r'_0 \in \arg\max\{\widehat{v}_i(\theta) : i \in [\![R']\!]\}$. In view of (30) ($r'$th row), we have

$$\frac{1}{R'} = \widehat{v}_{r'_0}(\theta) \sum_{r\in[\![R]\!]} (K(\theta))_{rr'_0}\widehat{u}_r(\theta) \geq k_\ell \|\widehat{v}(\theta)\|_\infty \|\widehat{u}(\theta)\|_\infty.$$

Hence, by (35),

$$\|\widehat{u}(\theta)\|_\infty \leq \frac{1}{k_\ell R' \|\widehat{v}(\theta)\|_\infty} \leq \frac{k_u R R'}{k_\ell R'} = \frac{k_u R}{k_\ell}. \tag{36}$$

Furthermore, for any $r' \in [\![R']\!]$, in view of (30) (r'th row) and (36),

$$\frac{1}{R'} = \widehat{v}_{r'}(\theta) \sum_{r\in[\![R]\!]} (K(\theta))_{rr'}\widehat{u}_r(\theta) \leq R k_u \|\widehat{u}(\theta)\|_\infty \widehat{v}_{r'}(\theta) \leq \frac{k_u^2 R^2}{k_\ell}\widehat{v}_{r'}(\theta). \tag{37}$$

The inequalities (35) and (37) readily imply (34). Likewise, for any $r \in [\![R]\!]$, in view of (29) (rth row),

$$\frac{1}{R} = \widehat{u}_r(\theta) \sum_{r'\in[\![R']\!]} (K(\theta))_{rr'}\widehat{v}_{r'}(\theta) \leq R' k_u \|\widehat{v}(\theta)\|_\infty \widehat{u}_r(\theta) \leq \frac{k_u R'}{k_\ell R}\widehat{u}_r(\theta). \tag{38}$$

The inequalities (36)and (38) readily imply (33).

- The function $\theta \mapsto \widehat{u}(\theta)$ is Lipschitz continuous on $\operatorname{dom} g_\tau$ 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 $\theta, \theta' \in \operatorname{dom} g_\tau$. Inequalities (33) and (34) imply that

$$\begin{aligned}\min\{(K(\theta)\widehat{v}(\theta'))_r : r \in [\![R]\!]\} &\geq k_\ell^2/(k_u^2 R^2),\\ \min\{(K(\theta)^\top \widehat{u}(\theta'))_{r'} : r' \in [\![R']\!]\} &\geq k_\ell^2 R/(k_u R').\end{aligned}$$

In view of Lemma 1 (first inequality), (22) (second inequality), (34) and the fact that K is L_K-Lipschitz (third inequality), we obtain

$$\begin{aligned} d_{\mathcal{H}}(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta)) &\leq \frac{2k_u^2 R^2}{k_\ell^2}\|K(\theta)\widehat{v}(\theta) - K(\theta')\widehat{v}(\theta)\|_2\\ &\leq \frac{2k_u^2 R^2}{k_\ell^2}\|K(\theta) - K(\theta')\|_F\|\widehat{v}(\theta)\|_2\\ &\leq \frac{2k_u^2 R\sqrt{R'}L_K}{k_\ell^3}\|\theta - \theta'\|_2. \end{aligned} \tag{39}$$

Likewise, using (33) instead of (34)

$$\begin{aligned} d_{\mathcal{H}}(K(\theta)^\top\widehat{u}(\theta), K(\theta')^\top\widehat{u}(\theta)) &\leq \frac{2k_u R'}{k_\ell^2 R}\|K(\theta_1)^\top\widehat{u}(\theta_1) - K(\theta_2)^\top\widehat{u}(\theta_1)\|_2\\ &\leq \frac{2k_u R'}{k_\ell^2 R}\|K(\theta)^\top - K(\theta')^\top\|_F\|\widehat{u}(\theta)\|_2\\ &\leq \frac{2k_u^2\sqrt{R}R' L_K}{k_\ell^3}\|\theta - \theta'\|_2. \end{aligned} \tag{40}$$

We can now bound the Hilbert projective metric between $\widehat{v}(\theta)$ and $\widehat{v}(\theta')$: by invoking in turn (31), (27), the triangle inequality, (28) and both (40) and (32), we get

$$\begin{aligned} d_{\mathcal{H}}(\widehat{v}(\theta), \widehat{v}(\theta')) &= d_{\mathcal{H}}\left(\frac{\mathbf{1}_{R'}/R'}{K(\theta)^\top\widehat{u}(\theta)}, \frac{\mathbf{1}_{R'}/R'}{K(\theta')^\top\widehat{u}(\theta')}\right)\\ &= d_{\mathcal{H}}\left(K(\theta)^\top\widehat{u}(\theta), K(\theta')^\top\widehat{u}(\theta')\right)\\ &\leq d_{\mathcal{H}}\left(K(\theta)^\top\widehat{u}(\theta), K(\theta')^\top\widehat{u}(\theta)\right) + d_{\mathcal{H}}\left(K(\theta')^\top\widehat{u}(\theta), K(\theta')^\top\widehat{u}(\theta')\right)\\ &\leq d_{\mathcal{H}}\left(K(\theta)^\top\widehat{u}(\theta), K(\theta')^\top\widehat{u}(\theta)\right) + \lambda\left(K(\theta')\right) d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta'))\\ &\leq \frac{2k_u^2\sqrt{R}R' L_K}{k_\ell^3}\|\theta - \theta'\|_2 + \Lambda d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta')). \end{aligned} \tag{41}$$

Likewise, by invoking in turn (31), (27), the triangle inequality, (28) and both (40) and (41), we get

$$\begin{aligned} d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta')) &= d_{\mathcal{H}}\left(\frac{\mathbf{1}_R/R}{K(\theta)\widehat{v}(\theta)}, \frac{\mathbf{1}_R/R}{K(\theta')\widehat{v}(\theta')}\right)\\ &= d_{\mathcal{H}}\left(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta')\right)\\ &\leq d_{\mathcal{H}}\left(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta)\right) + d_{\mathcal{H}}\left(K(\theta')\widehat{v}(\theta), K(\theta')\widehat{v}(\theta')\right) \end{aligned}$$

$$\leq d_{\mathcal{H}}(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta)) + \lambda(K(\theta'))d_{\mathcal{H}}(\widehat{v}(\theta), \widehat{v}(\theta'))$$
$$\leq d_{\mathcal{H}}(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta))$$
$$+ \Lambda\left(\frac{2k_u^2\sqrt{R}R'L_K}{k_\ell^3}\|\theta-\theta'\|_2 + \Lambda d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta'))\right).$$

The above inequality and (39) then yield

$$d_{\mathcal{H}}(\widehat{u}(\theta), \widehat{u}(\theta')) \leq \frac{1}{1-\Lambda^2}\left(d_{\mathcal{H}}(K(\theta)\widehat{v}(\theta), K(\theta')\widehat{v}(\theta)) + \Lambda\frac{2k_u^2\sqrt{R}R'L_K}{k_\ell^3}\|\theta-\theta'\|_2\right)$$
$$\leq \frac{2k_u^2\sqrt{RR'}L_K}{(1-\Lambda^2)k_\ell^3}(\sqrt{R}+\Lambda\sqrt{R'})\|\theta-\theta'\|_2.$$

Therefore, by Lemma 1 and (33), $\|\widehat{u}(\theta)-\widehat{u}(\theta')\|_2 \leq L_{\hat{u}}\|\theta-\theta'\|_2$, which completes the proof.

- The function $\theta \mapsto \widehat{v}(\theta)$ is Lipschitz continuous on $\operatorname{dom} g_\tau$ with Lipschitz constant

$$L_{\widehat{v}} := \frac{k_u L_K}{k_\ell^3\sqrt{R}} + \frac{k_u\sqrt{R'}L_{\hat{u}}}{k_\ell^2 R^{3/2}}.$$

Proof: Set arbitrarily $\theta, \theta' \in \operatorname{dom} g_\tau$. By (31) and (34),

$$\|\widehat{v}(\theta)-\widehat{v}(\theta')\|_2 = \left\|\frac{\mathbf{1}_{R'}/R'}{K(\theta)^\top\widehat{u}(\theta)} - \frac{\mathbf{1}_{R'}/R'}{K(\theta')^\top\widehat{u}(\theta')}\right\|_2$$
$$\leq \frac{\|K(\theta)^\top\widehat{u}(\theta) - K(\theta')^\top\widehat{u}(\theta')\|_2}{R'\min_{r'\in[\![R']\!]}\{(K(\theta_1)^\top\widehat{u}(\theta_1))_{r'}\}\min_{r'\in[\![R']\!]}\{(K(\theta')^\top\widehat{u}(\theta'))_{r'}\}}$$
$$= \frac{\|K(\theta)^\top\widehat{u}(\theta) - K(\theta')^\top\widehat{u}(\theta')\|_2}{\min_{r'\in[\![R']\!]}\{\widehat{v}_{r'}(\theta)^{-1}\}\min_{r'\in[\![R']\!]}\{\widehat{v}_{r'}(\theta')^{-1}\}}$$
$$\leq \frac{1}{k_\ell^2 R^2}\|K(\theta)^\top\widehat{u}(\theta) - K(\theta')^\top\widehat{u}(\theta')\|_2.$$

Moreover, using in turn the triangle inequality, (22) then the fact that K and $\widehat{u}$ are Lipschitz continuous and bounded on $\operatorname{dom} g_\tau$, we get

$$\|K(\theta)^\top\widehat{u}(\theta) - K(\theta')^\top\widehat{u}(\theta')\|_2 \leq \|K(\theta)^\top\widehat{u}(\theta) - K(\theta')^\top\widehat{u}(\theta)\|_2 + \|K(\theta')^\top\widehat{u}(\theta) - K(\theta')^\top\widehat{u}(\theta')\|_2$$
$$\leq \|K(\theta)-K(\theta')\|_F\|\widehat{u}(\theta)\|_2 + \|K(\theta')\|_F\|\widehat{u}(\theta)-\widehat{u}(\theta')\|_2$$
$$\leq \left(\frac{k_u R^{3/2}L_K}{k_\ell} + \sqrt{RR'}k_u L_{\hat{u}}\right)\|\theta-\theta'\|_2.$$

Therefore, $\|\widehat{v}(\theta)-\widehat{v}(\theta')\|_2 \leq L_{\hat{v}}\|\theta-\theta'\|_2$, which completes the proof.

- The function $\hat{P}(\theta)$ is Lipschitz continuous on $\operatorname{dom} g_\tau$.
Proof: We have proved that $\theta \mapsto \widehat{u}$, $\theta \mapsto K(\theta)$ and $\theta \mapsto \widehat{v}(\theta)$ are bounded and Lipschitz continuous on $\operatorname{dom} g_\tau$. Consequently, so is $\theta \mapsto \hat{P}(\theta) = \operatorname{diag}(\widehat{u}(\theta))K(\theta)\operatorname{diag}(\widehat{v}(\theta))$.

This completes the proof of Lemma 2, hence that of the fact that $\theta \mapsto \widehat{P}_\theta$ and $\theta \mapsto \widehat{Q}_\theta$ are Lipschitz continuous on $\operatorname{dom} g_\tau$.

A.1.4 The gradient of f is Lipschitz continuous

Set arbitrarily $\theta, \theta' \in \operatorname{dom} g_\tau \subset [0,1]^N$. We begin by noting that, by the triangle inequality and (22),

$$\begin{aligned}\tfrac{1}{2}\|\nabla f(\theta) - \nabla f(\theta')\|_2 &\leq \|y\|_2 \times \|\widehat{P}_\theta - \widehat{P}_{\theta'}\|_F + \|\theta\|_2 \times \|\widehat{Q}_\theta - \widehat{Q}_{\theta'}\|_F + \|\widehat{Q}_{\theta'}\|_F \times \|\theta - \theta'\|_2 \\ &\leq \|y\|_2 \times \|\widehat{P}_\theta - \widehat{P}_{\theta'}\|_F + \sqrt{N} \times \|\widehat{Q}_\theta - \widehat{Q}_{\theta'}\|_F + \|\theta - \theta'\|_2.\end{aligned}$$

We then readily conclude because we showed in Section A.1.3 that $\theta \mapsto \widehat{P}_\theta$ and $\theta \mapsto \widehat{Q}_\theta$ are Lipschitz continuous on $\operatorname{dom} g_\tau$.

A.2 The function H_δ satisfies the Kurdyka-Lojasiewicz property

A.2.1 The Kurdyka-Lojasiewicz property

Let us first recall what is the Kurdyka-Lojasiewicz property. Let $\ell : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ 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).

Definition 1 (Kurdyka-Lojasiewicz property, definition 3.1 in Attouch et al. (2010))**.** *The function ℓ is said to have the Kurdyka-Lojasiewicz property at $\bar{x} \in \operatorname{dom} \partial\ell$ if there exists $\eta \in (0, +\infty]$, a neighborhoood U of $\bar{x}$ and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ 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:*

$$\varphi'(\ell(x) - \ell(\bar{x}))\operatorname{dist}(0, \partial\ell(x)) \geq 1. \tag{42}$$

Inequality (42) 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 (42) 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 2 (Theorem 4.1 in Attouch et al. (2010))**.** *Any proper lower semicontinuous function $\ell : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ which is definable in an o-minimal structure $\mathcal{O}$ over $\mathbb{R}$ has the Kurdyka-Lojasiewicz property at each point of* $\operatorname{dom} \partial\ell$*. Moreover the function φ appearing in* (42) *is definable in $\mathcal{O}$.*

On *o*-minimal structures. An *o*-minimal structure over $\mathbb{R}$ can be viewed as an axiomatization of the quantitative properties of semialgebric sets. Semialgebric 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 $Q, R : \mathbb{R}^d \to \mathbb{R}$. Algebraic sets are finite unions and intersections of sets of the form $\{x \in \mathbb{R}^d : Q(x) = 0\}$ for some polynomial function $Q : \mathbb{R}^d \to \mathbb{R}$.

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 $p : \mathbb{R}^{n+1} \to \mathbb{R}^n$ 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 semialgebric sets is denoted $\mathbb{R}_{\text{alg}}$. It is the collection $\{\mathcal{O}_n\}_{n\geq 0}$ where each $\mathcal{O}_n$ is the class of semialgebric 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 $\exp : \mathbb{R} \to \mathbb{R}_+^*$ is denoted $\mathbb{R}_{\text{exp}}$. It extends $\mathbb{R}_{\text{alg}}$ and it is *o*-minimal over $\mathbb{R}$ (Wilkie, 1996).

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 $\varphi : A \to B$ between to definable sets is definable in $\mathcal{O}$ if its graph is definable in $\mathcal{O}$.

For instance, a polynomial function $Q : \mathbb{R}^d \to \mathbb{R}$ is definable in $\mathbb{R}_{\text{alg}}$, hence in $\mathbb{R}_{\text{exp}}$ 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 $\varphi : A \to B$ 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.

A.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 semialgebric hence definable.
  Proof: Introduce the sets $A_{r,r'} := \{P \in \mathbb{R}^{R\times R'} : P_{r,r'} \geq 0\}$, $B_r := \{P \in \mathbb{R}^{R\times R'} : \sum_{r'\in[\![R']\!]} P_{r,r'} = \frac{1}{R}\}$ and $C_{r'} := \{P \in \mathbb{R}^{R\times R'} : \sum_{r\in[\![R]\!]} P_{r,r'} = \frac{1}{R'}\}$ (for all $r \in [\![R]\!]$ and $r' \in [\![R']\!]$). Each of them is semialgebraic. Therefore their intersection, which equals $\Pi_{R,R'}$, is semialgebraic too, hence definable.

- Consider $F : \mathbb{R}^N \times \mathbb{R}^{M\times N} \times \mathbb{R}^{N\times N} \to \mathbb{R}$ given by

$$F(\theta, P, Q) := \sum_{m\in[\![M]\!],n\in[\![N]\!]} P_{m,n}\left(d(x_m, x'_n)^2 + (y_m - \theta_n)^2\right) - \tfrac{1}{2}\sum_{m\in[\![M]\!],n\in[\![N]\!]} Q_{n,n'}\left(d(x'_n, x'_{n'})^2 + (\theta_n - \theta_{n'})^2\right) + g_\tau(\theta).$$

  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, $\theta \mapsto \|\theta\|_1 = \sum_{n\in[\![N]\!]} |\theta_n|$ 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

$$\bigcup_{\varepsilon\in\{\pm1\}^N}\left[\bigcap_{n\in[\![N]\!]}\left\{\theta \in \mathbb{R}^N : \varepsilon_n\theta_n \geq 0\right\} \cap \left\{\theta \in \mathbb{R}^N : \sum_{n\in[\![N]\!]} \varepsilon_n\theta_n - \tau \leq 0\right\}\right],$$

    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 $\log : \mathbb{R}_+^* \to \mathbb{R}$ is definable (property (k)). Consequently, $\varphi : \mathbb{R}_+^* \to \mathbb{R}^2$ 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 $\Gamma_{\log}$ 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 $Q : \mathbb{R}^2 \to \mathbb{R}$ given by $Q(x, y) := -\gamma x(y - 1)$ is definable. Therefore, $\phi := Q \circ \varphi : \mathbb{R}_+^* \to \mathbb{R}$, 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 $\phi : \mathbb{R}_+ \to \mathbb{R}$. It follows that

$\gamma\mathcal{E} : (\mathbb{R}_+)^{R\times R'} \to \mathbb{R}$ given by $\gamma\mathcal{E}(P) := \sum_{r\in[\![R]\!],r'\in[\![R']\!]} \phi(P_{r,r'})$ is definable (property (i)), hence $\gamma E := \gamma\mathcal{E}|_{\Pi_{R,R'}}$ is definable too (property (g)).

- The function $(f + g_\tau) : \mathbb{R}^N \to \mathbb{R}$ is definable.
  Proof: This is a straightforward consequence of the fact that, for all $\theta \in \mathbb{R}^N$,

$$(f + g_\tau)(\theta) := \min_{P\in\Pi_{M,N}} \max_{Q\in\Pi_{N,N}} \left\{ F|_{\mathbb{R}^N\times\Pi_{M,N}\times\Pi_{N,N}} + \gamma E(P) - \tfrac{1}{2}\gamma E(Q) \right\},$$

  where the sets $\Pi_{M,N}$ and $\Pi_{N,N}$ are definable (property (m)).

- The function H_δ is definable.
  Proof: Recall that $H_\delta : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ 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)).

A.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 A.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 2, H_δ satisfies the Kurdyka-Lojasiewicz property on the domain of ∂H_δ.