Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Computational Intelligence & Inform. Management

On sparse ensemble methods: An application to short-term predictions of the evolution of COVID-19

Sandra Benítez-Peña[a,b], Emilio Carrizosa[a,b], Vanesa Guerrero[c], M. Dolores Jiménez-Gamero[a,b], Belén Martín-Barragán[d], Cristina Molero-Río[a,b], Pepa Ramírez-Cobo[e,a], Dolores Romero Morales[f,*], M. Remedios Sillero-Denamiel[a,b]

[a] *Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain*
[b] *Departamento de Estadística e Investigación Operativa, Universidad de Sevilla, Seville, Spain*
[c] *Departamento de Estadística, Universidad Carlos III de Madrid, Getafe, Spain*
[d] *The University of Edinburgh Business School, University of Edinburgh, Edinburgh, UK*
[e] *Departamento de Estadística e Investigación Operativa, Universidad de Cádiz, Cadiz, Spain*
[f] *Department of Economics, Copenhagen Business School, Frederiksberg, Denmark*

ARTICLE INFO

Article history:
Received 27 May 2020
Accepted 7 April 2021
Available online 18 April 2021

Keywords:
Machine Learning
Ensemble Method
Mathematical Optimization
Selective Sparsity
COVID-19

ABSTRACT

Since the seminal paper by Bates and Granger in 1969, a vast number of ensemble methods that combine different base regressors to generate a unique one have been proposed in the literature. The so-obtained regressor method may have better accuracy than its components, but at the same time it may overfit, it may be distorted by base regressors with low accuracy, and it may be too complex to understand and explain. This paper proposes and studies a novel Mathematical Optimization model to build a sparse ensemble, which trades off the accuracy of the ensemble and the number of base regressors used. The latter is controlled by means of a regularization term that penalizes regressors with a poor individual performance. Our approach is flexible to incorporate desirable properties one may have on the ensemble, such as controlling the performance of the ensemble in critical groups of records, or the costs associated with the base regressors involved in the ensemble. We illustrate our approach with real data sets arising in the COVID-19 context.

1. Introduction

A plethora of methodologies of very different nature is currently available for predicting a continuous response variable, as it is the case in regression as well as in time series forecasting. Those methodologies come mainly from Machine Learning, such as Support Vector Machines (Carrizosa & Romero Morales, 2013; Vapnik, 1995), Random Forests (Breiman, 2001), Optimal Trees (Bertsimas & Dunn, 2017; Blanquero, Carrizosa, Molero-Río, & Romero Morales, 2021; Carrizosa, Molero-Río, & Romero Morales, 2021), Deep Learning (Gambella, Ghaddar, & Naoum-Sawaya, 2021); or from Statistics, such as Generalized Linear Models (Hastie, Tibshirani, & Wainwright, 2015), Semi- and Nonparametric approaches to regression (such as smoothing techniques) (Härdle, 1990), Regression models for time series analysis (Kedem & Fokianos, 2005), or Random Effects models (Lee, Nelder, & Pawitan, 2018). Some of these techniques have shown a relatively high degree of success in COVID-19 time series forecasting (Benítez-Peña et al., 2020b; Nikolopoulos, Punia, Schäfers, Tsinopoulos, & Vasilakis, 2021), which is the application that has inspired this work.

In this way, the user has at hand a long list of fitted regression models, referred to in what follows as base regressors, and faces the problem of deciding which one to choose, or alternatively, how to combine (some of) the competing approaches, that is, how to build an ensemble. While a thorough computational study of the different models may help the user to identify the most convenient one, such an approach becomes unworkable when predicting

[*] Corresponding author.

E-mail addresses: sbenitez1@us.es (S. Benítez-Peña), ecarrizosa@us.es (E. Carrizosa), vanesa.guerrero@uc3m.es (V. Guerrero), dolores@us.es (M.D. Jiménez-Gamero), Belen.Martin@ed.ac.uk (B. Martín-Barragán), mmolero@us.es (C. Molero-Río), pepa.ramirez@uca.es (P. Ramírez-Cobo), drm.eco@cbs.dk (D. Romero Morales), rsillero@us.es (M.R. Sillero-Denamiel).

https://doi.org/10.1016/j.ejor.2021.04.016
0377-2217/

new phenomena in real-time, like the evolution of the COVID-19 counts (confirmed cases, hospitalized patients, ICU patients, recovered patients, and fatalities). Here, the most accurate method will probably change over time since we are dealing with a dynamic setting, but also because of the non-stationarity of the data caused, for instance, by the different interventions of authorities to *flatten the curve*.

Hence, it may be more convenient to build an ensemble where some accuracy measure, such as a (cross-validation) estimate of the expected squared error or of the absolute error (Ando & Li, 2014; Bates & Granger, 1969), is optimized at each forecast origin. With this approach other relevant issues can be modeled, such as sparsity in the feature space (Bertsimas, King, & Mazumder, 2016; Carrizosa, Mortensen, Romero Morales, & Sillero-Denamiel, 2020a; Carrizosa, Olivares-Nadal, & Ramírez-Cobo, 2017b; Fountoulakis & Gondzio, 2016), interpretability (Carrizosa, Nogales-Gómez, & Romero Morales, 2016; 2017a; Carrizosa, Olivares-Nadal, & Ramírez-Cobo, 2020b; Martín-Barragán, Lillo, & Romo, 2014), critical values of features (Carrizosa, Martín-Barragán, & Romero Morales, 2010; 2011), measurement costs (Carrizosa, Martín-Barragán, & Romero Morales, 2008), or cost-sensitive performance constraints (Benítez-Peña, Blanquero, Carrizosa, & Ramírez-Cobo, 2019a; 2020a; Blanquero, Carrizosa, Ramírez-Cobo, & Sillero-Denamiel, 2020). See (Friese, Bartz-Beielstein, & Emmerich, 2016; Mendes-Moreira, Soares, Jorge, & Sousa, 2012; Ren, Zhang, & Suganthan, 2016) and references therein for the role of mathematical optimization when constructing ensembles and (Friese, Bartz-Beielstein, Bäck, Naujoks, & Emmerich, 2019) for the use of ensembles to enhance the optimization of black-box expensive functions.

In this paper, we propose an optimization approach to build a sparse ensemble. In contrast to existing proposals in the literature, our paper focuses on an innovative definition of sparsity, the so-called *selective sparsity*. Our goal is to build a sparse ensemble, which takes into account the individual performance of each base regressor, in such a way that only *good base regressors* are allowed to take part in the ensemble. This is done with the aim to adapt to dynamic settings, such as in COVID-19 counts, where the composition of the ensemble may change over time, but also to avoid that the ensemble is distorted by base regressors with low accuracy or may be too complex to understand and explain. Ours can be seen as a sort of what (Mendes-Moreira et al., 2012) calls an *ensemble pruning*, where the ensemble is constructed by using a subset of all available base regressors. The novelty of our approach resides in the fact that the selection of the subset and the weights in the ensemble are simultaneously optimized.

We propose a Mathematical Optimization model that trades off the accuracy of the ensemble and the number of base regressors used. The latter is controlled by means of a regularization term that penalizes regressors with a poor individual performance. Our approach is flexible to incorporate desirable properties one may have on the ensemble, such as controlling the performance of the ensemble in critical groups of records, or the costs associated with the base regressors involved in the ensemble. Our data-driven approach is applied to short-term predictions of the evolution of COVID-19, as an alternative to model-based prediction algorithms as in Achterberg et al. (2020) and references therein.

The remainder of the paper is structured as follows. Section 2 formulates the Mathematical Optimization problem to construct the sparse ensemble. Theoretical properties of the optimal solution are studied, and how to accommodate some desirable properties on the ensemble is also discussed. Section 3 illustrates our approach with real data sets arising in the COVID-19 context, where one can see how the ensemble composition changes over time. The paper ends with some concluding remarks and lines for future research in Section 4.

2. The optimization model

This section presents the new ensemble approach. Section 2.1 describes the formulation of the model in terms of an optimization problem with linear constraints. Section 2.2 establishes the connection of the approach with the constrained Lasso (Blanquero et al., 2020; Gaines, Kim, & Zhou, 2018) and some theoretical results of the solution are derived. Finally, Section 2.3 considers some extensions of the model concerning the control of the set of base regressors or control of the performance in critical groups.

2.1. The formulation

Let $\mathcal{F}$ be a finite set of base regressors for the response variable y. No restriction is imposed on the collection of base regressors. It may include a variety of state-of-the-art models and methodologies for setting their parameters and hyperparameters. It may even use alternative samples for training, for example where individuals are characterized by different sets of features. By taking convex combinations of the base regressors in $\mathcal{F}$, we obtain a broader class of regressors, namely, $co(\mathcal{F}) = \left\{F = \sum_{f\in\mathcal{F}} \alpha_f f : \sum_{f\in\mathcal{F}} \alpha_f = 1, \alpha_f \geq 0, \ \ \forall f \in \mathcal{F}\right\}$. Throughout this section, vectors will be denoted with bold typesetting, e.g., $\boldsymbol{\alpha} = (\alpha_f)_{f\in\mathcal{F}}$.

The selection of one combined regressor from $co(\mathcal{F})$ will be made by optimizing a function which takes into account two criteria. The first and fundamental criterion is the overall accuracy of the combined regressor, measured through a loss function $\mathcal{L}$, defined on $co(\mathcal{F})$,

$$\begin{aligned} \mathcal{L} : \ co(\mathcal{F}) &\longmapsto \mathbb{R} \\ F &\longmapsto \mathcal{L}(F). \end{aligned}$$

For each base regressor $f \in \mathcal{F}$ we assume its individual loss $\mathcal{L}_f$ is given. This may be simply defined as $\mathcal{L}_f = \mathcal{L}(f)$, but other options are possible too, in which, for instance, $\mathcal{L}_f$ and $\mathcal{L}$ are both empirical losses, as in Section 2.2, but use different training samples.

With the second criterion, a selective sparsity is pursued to make the method more reluctant to choose base regressors $f \in \mathcal{F}$ with lower reliability, i.e., with higher individual loss $\mathcal{L}_f$, reducing thus overfitting. To achieve this, we add a regularization term in which the weight of base regressor f, say α_f, is multiplied by its individual loss $\mathcal{L}_f$. The selective sparse ensemble is obtained by solving the following Mathematical Optimization problem with linear constraints:

$$\min_{\boldsymbol{\alpha}\in\mathcal{S}} \left\{ \mathcal{L}\left(\sum_{f\in\mathcal{F}} \alpha_f f \right) + \lambda \sum_{f\in\mathcal{F}} \alpha_f \mathcal{L}_f \right\}, \tag{1}$$

where $\mathcal{S}$ is the unit simplex in $\mathbb{R}^{|\mathcal{F}|}$,

$$\mathcal{S} = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{F}|} : \sum_{f\in\mathcal{F}} \alpha_f = 1, \alpha_f \geq 0, \ \ \forall f \in \mathcal{F} \right\},$$

and $\lambda \geq 0$ is a regularization parameter, which trades off the importance given to the loss of the ensemble regressor and to the selective sparsity of the base regressors used.

2.2. Theoretical results

In general, Problem (1) has a nonlinear objective function and linear constraints. For loss functions commonly used in the literature, we can rewrite its objective as a linear or a convex quadratic function while the constraints remain linear. Therefore, for these loss functions, Problem (1) is easily tractable with commercial solvers. In addition, and under some mild assumptions, we characterize the behavior of the optimal solution with respect to the parameter λ.

First, we will rewrite the second term in the objective function, so that the proposed model can be seen as a particular case of the constrained Lasso. As for Lasso models and extensions of them, having a sparse model reduces the danger of overfitting.

Remark 1. The so-called selective ℓ_1 norm $\|\cdot\|_1^{\text{sel}}$ in $\mathbb{R}^{|\mathcal{F}|}$ is defined as

$$\|\boldsymbol{\alpha}\|_1^{\text{sel}} = \sum_{f\in\mathcal{F}} \mathcal{L}_f|\alpha_f|.$$

The objective function in Problem (1) can be written as $\mathcal{L}\big(\sum_{f\in\mathcal{F}}\alpha_f f\big) + \lambda\|\boldsymbol{\alpha}\|_1^{\text{sel}}$. With this, and for well-known losses $\mathcal{L}$, Problem (1) can be seen as a constrained Lasso problem, (Blanquero et al., 2020; Gaines, Kim, & Zhou, 2018), in which a selective sparsity is sought, as opposed to a plain sparsity with as few nonzero coefficients α_f as possible. □

Remark 2. Let $\mathcal{I}$ be a training sample, in which each individual $i \in \mathcal{I}$ is characterized by its feature vector $\mathbf{x}_i \in \mathbb{R}^p$ and its response y_i. Let $\mathcal{L}$ be the empirical loss of quantile regression, (Koenker & Hallock, 2001), for $\mathcal{I}$,

$$\mathcal{L}\left(\sum_{f\in\mathcal{F}}\alpha_f f\right) = \sum_{i\in\mathcal{I}}\rho_\tau\left(y_i - \sum_{f\in\mathcal{F}}\alpha_f f(\mathbf{x}_i)\right), \tag{2}$$

where

$$\rho_\tau(s) = \begin{cases} \tau s, & \text{if } s \geq 0 \\ -(1-\tau)s, & \text{if } s < 0, \end{cases}$$

for some $\tau \in (0,1)$. Then, as in e.g. Koenker and Ng (2005), Problem (1) can be expressed as a linear program and thus efficiently solved with Linear Programming solvers. □

Remark 3. Let $\mathcal{I}$ be a training sample, in which each individual $i \in \mathcal{I}$ is characterized by its feature vector $\mathbf{x}_i \in \mathbb{R}^p$ and its response y_i. Let $\mathcal{L}$ be the empirical loss of Ordinary Least Squares (OLS) regression for $\mathcal{I}$, i.e.,

$$\mathcal{L}\left(\sum_{f\in\mathcal{F}}\alpha_f f\right) = \sum_{i\in\mathcal{I}}\left(y_i - \sum_{f\in\mathcal{F}}\alpha_f f(\mathbf{x}_i)\right)^2. \tag{3}$$

Hence, Problem (1) is a convex quadratic problem with linear constraints, which, by Remark 1, can be seen as a constrained Lasso. In particular, the results in Gaines, Kim, and Zhou, (2018) apply, and thus, we can assert that, if the design matrix $(f(\mathbf{x}_i))_{i\in\mathcal{I}, f\in\mathcal{F}}$ has full rank, then,

1. For any $\lambda \geq 0$, Problem (1) has unique optimal solution $\boldsymbol{\alpha}^\lambda$.
2. The path of optimal solutions $\boldsymbol{\alpha}^\lambda$ is piecewise linear in λ. □

Under mild conditions on $\mathcal{L}$, applicable in particular for the quantile and OLS empirical loss functions, we characterize the optimal solution of Problem (1) for large values of the parameter λ. Intuitively speaking, for λ growing to infinity, the first term in the objective function becomes negligible, and thus we only need to solve the Linear Programming problem of minimizing $\sum_{f\in\mathcal{F}}\alpha_f\mathcal{L}_f$ in the simplex $\mathcal{S}$. This problem attains its optimum at one of the extreme points of the feasible region, i.e., at some $f^* \in \mathcal{F}$, namely, one for which $\mathcal{L}_{f^*} \leq \mathcal{L}_f$, $\forall f$. We formalize this intuition in the following proposition, where under the assumption of convexity of $\mathcal{L}$, we show that a finite value of λ exists for which such sparse solution is optimal. Before stating it, notice that, since the set $\mathcal{F}$ is given, we can define

$$\begin{aligned} L:\ \Omega &\longmapsto \mathbb{R} \\ \mathbf{w} &\longmapsto L(\mathbf{w}) = \mathcal{L}\left(\sum_{f\in\mathcal{F}} w_f f\right), \end{aligned}$$

for some $\Omega \subseteq \mathbb{R}^{|\mathcal{F}|}$, such that $\Omega \supseteq \mathcal{S}$.

Proposition 1. *Assume that L is convex in an open convex set $\Omega \supseteq \mathcal{S}$. Furthermore, assume that there exists a base regressor f° such that $\mathcal{L}_{f^\circ} < \mathcal{L}_f$ for all $f \in \mathcal{F}$, $f \neq f^\circ$. Then, there exists $\lambda^\circ < +\infty$ such that, for any $\lambda \geq \lambda^\circ$, f° is an optimal solution to Problem (1).*

Proof. Let f° be as in the statement of the proposition, and let $\boldsymbol{\alpha}^\circ \in \mathcal{S}$ denote the vector with 1 in its component corresponding to f° and 0 otherwise. Since L is defined in the open set $\Omega \ni \boldsymbol{\alpha}^\circ$, the subdifferential $\partial L(\boldsymbol{\alpha}^\circ)$ of the convex function L at $\boldsymbol{\alpha}^\circ$ is not empty. Let $\boldsymbol{p} \in \partial L(\boldsymbol{\alpha}^\circ)$, and let $\mathcal{N}(\boldsymbol{\alpha}^\circ)$ denote the normal cone of $\mathcal{S}$ at $\boldsymbol{\alpha}^\circ$. Then,

$$0 \in \boldsymbol{p} + \lambda\big(\mathcal{L}_f\big)_{f\in\mathcal{F}} + \mathcal{N}(\boldsymbol{\alpha}^\circ) \text{ iff } p_{f^\circ} + \lambda\mathcal{L}_{f^\circ} \leq p_f + \lambda\mathcal{L}_f \qquad \forall f \in \mathcal{F}, \tag{4}$$

which is satisfied iff

$$\lambda \geq \max\left\{\frac{p_{f^\circ} - p_f}{\mathcal{L}_f - \mathcal{L}_{f^\circ}} : f \in \mathcal{F}, f \neq f^\circ\right\}. \tag{5}$$

Setting λ° equal to the value on the right-hand side of (5), and taking into account that the condition on the left-hand side of (4) is necessary and sufficient for the optimality of $\boldsymbol{\alpha}^\circ$, the result follows. □

2.3. Extensions

Problem (1) can be enriched to address some desirable properties one may seek for the ensemble. Three of them are discussed in what follows. The first two properties relate to the transparency and interpretability of the ensemble, Deng (2019) and Florez-Lopez and Ramon-Jeronimo (2015), while the third one relates to the performance of the ensemble in critical groups.

As mentioned in the introduction, the ensemble may contain base regressors built with several methodologies of very diverse nature. Therefore, one may want to control the number of methodologies used in the final ensemble. For instance, in the application described in Section 3, we consider four methodologies, namely, Support Vector Regression, Random Forests, Optimal Trees, and Logistic Regression. Let $\mathcal{F} = \bigcup_{m\in\mathcal{M}}\mathcal{F}_m^{\text{type}}$, where $\mathcal{F}_m^{\text{type}}$ is the set of base regressors using methodology $m \in \mathcal{M}$, and let $\boldsymbol{\alpha}_m^{\text{type}}$ be the corresponding subvector of $\boldsymbol{\alpha}$, namely, the one containing the components in $\boldsymbol{\alpha}$ referring to methodology $m \in \mathcal{M}$. With this, we can extend the objective function of Problem (1) to

$$\mathcal{L}\left(\sum_{f\in\mathcal{F}}\alpha_f f\right) + \lambda\sum_{f\in\mathcal{F}}\alpha_f\mathcal{L}_f + \lambda^{\text{type}}\sum_{m\in\mathcal{M}}\|\boldsymbol{\alpha}_m^{\text{type}}\|_\infty. \tag{6}$$

In a similar fashion, one may want to control the set of features used by the ensemble. Let $\mathcal{F}_j^{\text{fea}} \subseteq \mathcal{F}$ be the set of base regressors using feature $j \in \{1,\ldots,p\}$, and let $\boldsymbol{\alpha}_j^{\text{fea}}$ be the corresponding subvector of $\boldsymbol{\alpha}$, namely, the one containing the components in $\boldsymbol{\alpha}$ referring to feature $j \in \{1,\ldots,p\}$. With this, we can extend the objective function of Problem (1) to

$$\mathcal{L}\left(\sum_{f\in\mathcal{F}}\alpha_f f\right) + \lambda\sum_{f\in\mathcal{F}}\alpha_f\mathcal{L}_f + \lambda^{\text{fea}}\sum_{j=1}^{p}\|\boldsymbol{\alpha}_j^{\text{fea}}\|_\infty. \tag{7}$$

In both cases, the ℓ_∞ terms can be rewritten using new decision variables and linear constraints, and thus the structure of the problem is not changed. This way, if $\mathcal{L}$ is the quantile regression

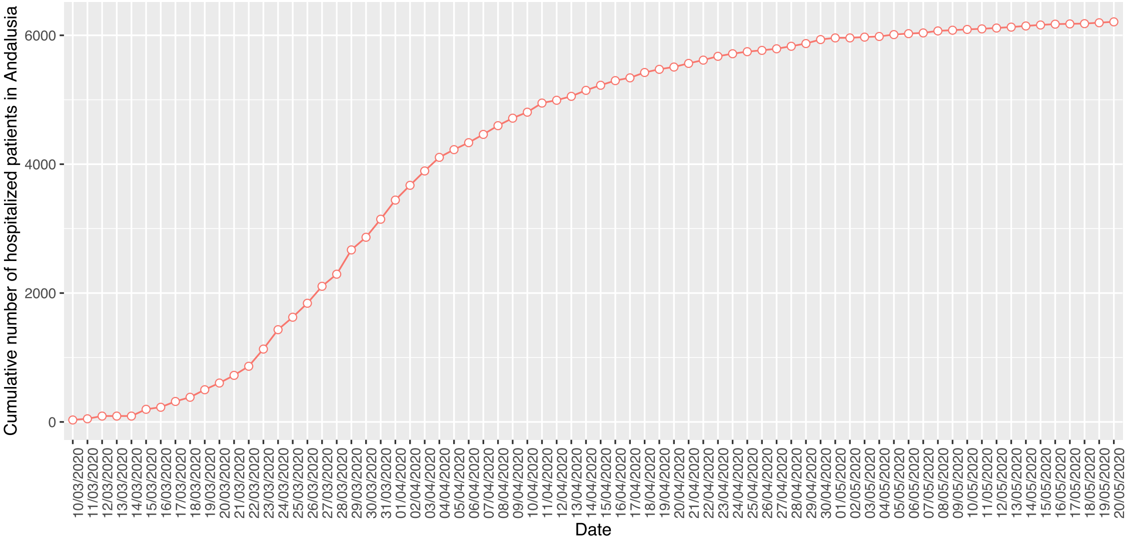


Fig. 1. Cumulative number of hospitalized patients in Andalusia (Spain) for COVID-19 in the period 10/03/2020–20/05/2020.

(respectively, the Ordinary Least Squares) empirical loss, the optimization problem with objective as in (6) is written as a linear problem (respectively, as a convex quadratic problem with linear constraints). The same holds for the optimization problem with objective as in (7).

In addition, our approach can easily incorporate cost-sensitive performance constraints to ensure that we control not only the overall accuracy of the regressor, but also the accuracy on a number of critical groups, as in Benítez-Peña et al. (2019a), Benítez-Peña, Blanquero, Carrizosa, and Ramírez-Cobo (2019b), Blanquero et al. (2020) and Datta and Das (2015). With this, if $\delta^g > 0$ denotes the threshold on the loss $\mathcal{L}^g$ for group $g \in \mathcal{G}$, we can add to the feasible region of Problem (1) constraints

$$\mathcal{L}^g\left(\sum_{f\in\mathcal{F}}\alpha_f\, f\right) \leq \delta^g,\ \forall g \in \mathcal{G}. \tag{8}$$

For the quantile and Ordinary Least Squares empirical loss functions, these constraints are linear or convex quadratic, respectively, and thus the optimization problems can be addressed with the very same numerical tools as before.

3. Short-term predictions of the evolution of COVID-19

The purpose of this section is to illustrate how, thanks to the selective sparsity term in Problem (1), we can provide good ensembles in terms of accuracy. For this, we use data sets arising in the context of COVID-19.

3.1. The data

COVID-19 was first identified in China in December 2019 and, subsequently, started to spread broadly. Quickly after this, data started to be collected daily by the different countries. Several variables of interest, such as confirmed cases, hospitalized patients, ICU patients, recovered patients, and fatalities, among others, were considered. Different initiatives around the world emerged in order to get to know this new scenario.

In this section, we focus on the evolution of the pandemic in Spain and Denmark. The first cases were confirmed in Spain and Denmark in late February 2020 and early March 2020, respectively. In this paper, the considered variable of interest is the cumulative number of hospitalized patients in the regions of Andalusia (Spain) and Sjælland (Denmark). Figs. 1 and 2 display the data in the periods 10/03/2020-20/05/2020 for Andalusia and 06/03/2020-20/05/2020 for Sjælland, which can be found at the repositories in Fernández-Casal (2020) and Statens Serum Institut (2020), respectively.

The univariate time series $\{X_t,\ t = 1,\dots,T\}$, with X_t representing the cumulative number of hospitalized patients in the region under consideration in day t, is converted into a multivariate series using seven lags. In other words, the data fed to the base regressors is not the time series itself, but the vectors of covariates and responses in Fig. 3. This training set is just one of the different options we have considered to create base regressors. In the next section, we discuss other data choices, which we will refer to as `Country`, `Transformation` and `Differences`.

3.2. Options for feeding the data

We first discuss the `Country` data choice. Let R be the number of regions of the country under consideration, and, without loss of generalization, let us assume that the first one is the region under consideration. The time series $\left\{X_t^r,\ t = 1,\dots,T\right\}$, for regions $r = 2,\dots,R$, were also available. Such times series are correlated with the one under consideration. We had to decide whether to incorporate these additional time series in our forecasting model. If we do so, the feeding data contains the 7-uples in Fig. 3 from the region under consideration, as well as the ones from the other $R-1$ regions, see Fig. 4. We now move to the `Transformation` choice. For the two choices in Figs. 3 and 4, either the crude data X are used or they are transformed using some standard Box-Cox transformations, Hastie, Tibshirani and Wainwright (2015), namely, X^2 and $\log(X+1)$. Finally, with respect to the `Differences` choice, we have also considered whether information about the monotonicity (first difference, $\Delta X_t := X_t - X_{t-1}$) and the curvature (second difference, $\Delta^2 X_t := \Delta X_t - \Delta X_{t-1}$) is added to the feeding data as predictors, thus yielding 6 and 5 new predictors because of monotonicity and curvature, respectively.

To end this section, observe that the time series $\{X_t,\ t = 1,\dots,T\}$ of cumulative number of hospitalized patients in the region under consideration is, by nature, nondecreasing. However, some of the methodologies in the next section used to build base regressors do not guarantee such monotonicity. To ensure that the predictions show the monotonicity property present in the data, we use as response variable $\log(1+\Delta X_t)$, instead of

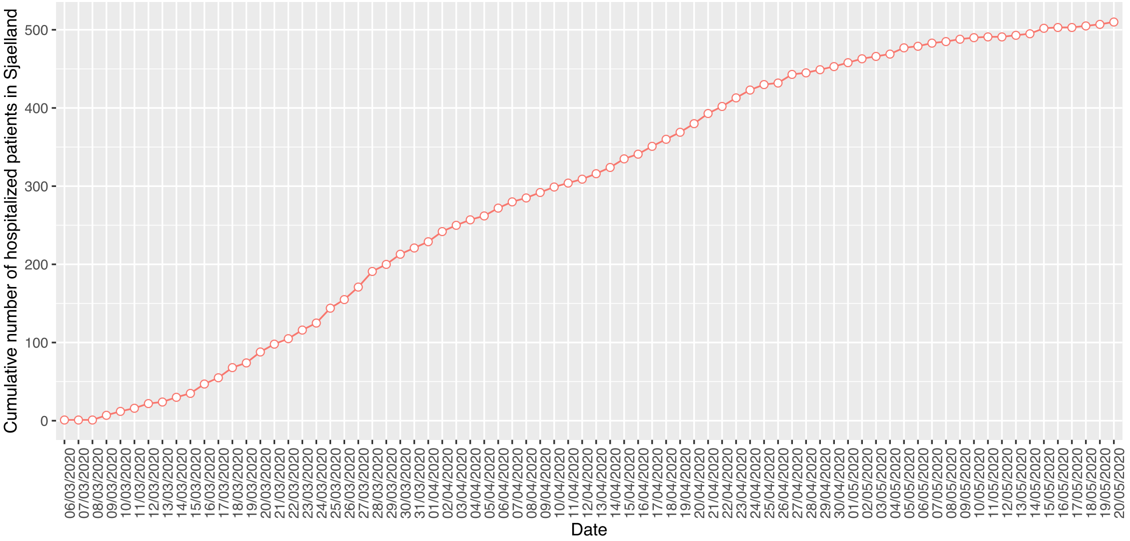


Fig. 2. Cumulative number of hospitalized patients in Sjælland (Denmark) for COVID-19 in the period 06/03/2020–20/05/2020.

$$\begin{array}{ccccc}
(X_1, & X_2, & \ldots & X_7) & X_8 \\
 & & \vdots & & \vdots \\
(X_{T-7}, & X_{T-6}, & \ldots & X_{T-1}) & X_T
\end{array}$$

Fig. 3. Covariates (in parentheses) and response variable for the cumulative number of hospitalized patients in the region under consideration.

$$\begin{array}{ccccc}
(X_1, & X_2, & \ldots & X_7) & X_8 \\
 & & \vdots & & \vdots \\
(X_{T-7}, & X_{T-6}, & \ldots & X_{T-1}) & X_T \\
(X_1^2, & X_2^2, & \ldots & X_7^2) & X_8^2 \\
 & & \vdots & & \vdots \\
(X_{T-7}^2, & X_{T-6}^2, & \ldots & X_{T-1}^2) & X_T^2 \\
 & & \vdots & & \vdots \\
(X_1^{\mathrm{R}}, & X_2^{\mathrm{R}}, & \ldots & X_7^{\mathrm{R}}) & X_8^{\mathrm{R}} \\
 & & \vdots & & \vdots \\
(X_{T-7}^{\mathrm{R}}, & X_{T-6}^{\mathrm{R}}, & \ldots & X_{T-1}^{\mathrm{R}}) & X_T^{\mathrm{R}}
\end{array}$$

Fig. 4. Covariates (in parentheses) and response variable for the cumulative number of hospitalized patients in each of the R regions of the country.

X_t. Once the procedure is completed, we undo this transformation to predict the original response variable X_t. Figs. 5 and 6 display $\log(1+\Delta X_t)$ for Andalusia and Sjælland, respectively, where t is as in Figs. 1 and 2.

3.3. The base regressors

We consider four base methodologies to build the set of base regressors $\mathcal{F}$. This includes three state-of-the-art Machine Learning tools, namely Support Vector Regression (SVR) (Carrizosa & Romero Morales, 2013), Random Forest (RF) (Breiman, 2001), and Sparse Optimal Randomized Regression Trees (S-ORRT) (Blanquero, Carrizosa, Molero-Río, & Romero Morales, 2020a), as well as the classic Linear Regression (LR). Each of them is fed each time with one of the data choices described in Section 3.1. See Table 1 for a description of the elements of $\mathcal{F} = \mathcal{F}_{\text{SVR}} \cup \mathcal{F}_{\text{RF}} \cup \mathcal{F}_{\text{LR}} \cup \mathcal{F}_{\text{S-ORRT}} = \left\{f_j : j = 1, \ldots, 36\right\}$ according to their methodology and the data choices. These methodologies have some parameters which must be tuned, and we explain below the tuning we have performed together with other computational details.

To tune the parameters, the different base regressors are trained using all the available data, except for the last four days, i.e., these models are trained on $t \in \{1, \ldots, T-4\}$. The `e1071` (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2019) and `randomForest` (Liaw & Wiener, 2002) R packages have been used for training SVR and RF, respectively, while the `lm` routine in R is used for LR. The computational details for training S-ORRT are those in Blanquero et al. (2020a). For SVR, we use the RBF kernel and perform a grid search in $\{2^a : a = -10, \ldots, 10\}$ for both parameters, `cost` and `gamma`. For RF, we set `ntree` $= 500$ and for `mtry` we try out five random values. If only information from the region under consideration is included ('`Country` No' data option in Table 1), eight fold cross-validation is used. However, when information from all regions in the country is included, we limit this to five fold cross-validation, due to the small amount of data and the lack of observations in some regions. Such cross-validation estimates are used to select the best values of the parameters. With those best values, for each combination of feeding data and methodology, the base regressors $f \in \mathcal{F}$ are built using information from $t \in \{1, \ldots, T-4\}$, see Fig. 7.

3.4. The pseudocode of the complete procedure

The complete procedure for making short-term predictions with our selective sparse ensemble methodology is summarized in Algorithm 1 and can be visualized in Fig. 7. The considered grid of values for the tradeoff parameter λ in Problem (1) is $\left\{0, 2^{-10}, 2^{-9}, \ldots, 2^3\right\}$. For the tests considered in this section, this grid is wide enough. On one extreme, we have included the trivial value $\lambda = 0$, for which the selective sparsity term does not play a role. On the other extreme, with this grid we ensure that $\lambda = \lambda^\circ$ is reached, for which, by Proposition 1, the ensemble shows the highest level of sparsity.

We start by training the base regressors $\mathcal{F}$ in Table 1, with tuning parameters as in Section 3.3, using the data available up to day $T-4$. We then move to solve Problem (1) for the different values of λ in the grid. For this, we have chosen the loss $\mathcal{L}$ as in (3), where $\mathcal{I}$ consists of the data in the four days left out when tuning

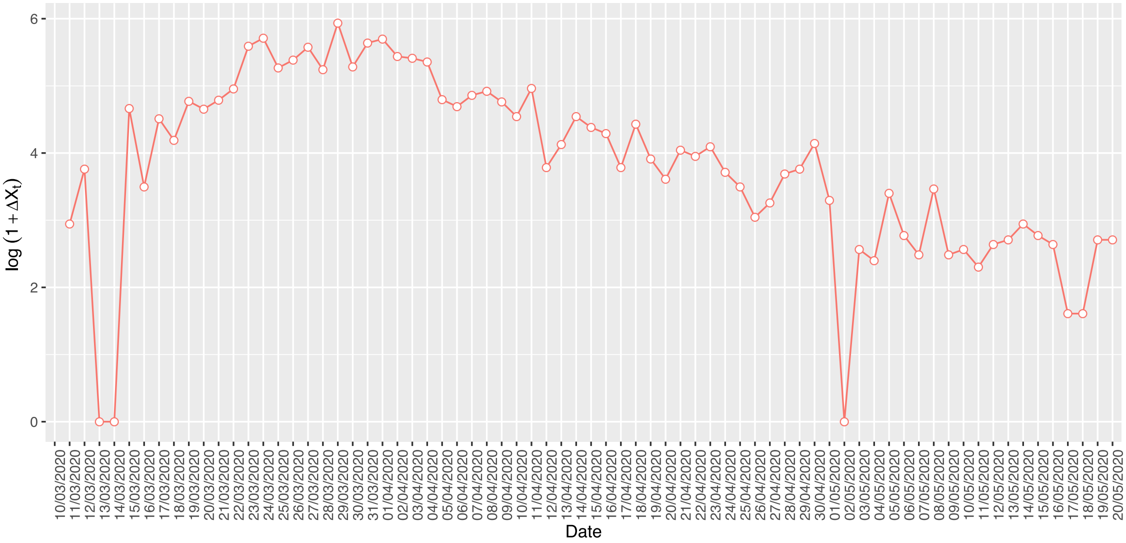


Fig. 5. Representation of the function $\log(1+\Delta X_t)$, where X_t denote the cumulative number of hospitalized patients in Andalusia for COVID-19 in the period 10/03/2020–20/05/2020.

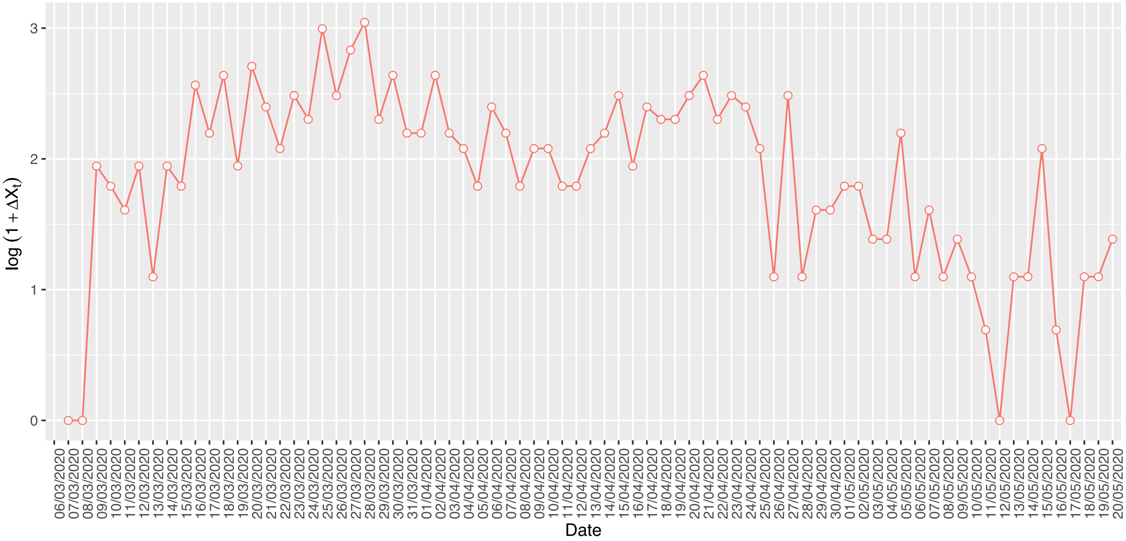


Fig. 6. Representation of the function $\log(1+\Delta X_t)$, where X_t denote the cumulative number of hospitalized patients in Sjælland for COVID-19 in the period 06/03/2020–20/05/2020.

Table 1
Description of the chosen base regressors according to the data choices on `Country`, `Transformation` and `Differences` and the four methodologies used, with tuning parameters as in Section 3.3.

	$\mathcal{F}_{\text{SVR}}$												$\mathcal{F}_{\text{RF}}$					
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}
`Country` No	✓	✓			✓	✓			✓	✓			✓		✓		✓	
`Country` Yes			✓	✓			✓	✓			✓	✓		✓		✓		✓
`Transformation` X	✓	✓	✓	✓									✓	✓				
`Transformation` $\log(X+1)$					✓	✓	✓	✓							✓	✓		
`Transformation` X^2									✓	✓	✓	✓					✓	✓
`Differences` Yes	✓		✓		✓		✓		✓		✓		✓	✓	✓	✓	✓	✓
`Differences` No		✓		✓		✓		✓		✓		✓						
	$\mathcal{F}_{\text{LR}}$						$\mathcal{F}_{\text{S-ORRT}}$											
	f_{19}	f_{20}	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}	f_{26}	f_{27}	f_{28}	f_{29}	f_{30}	f_{31}	f_{32}	f_{33}	f_{34}	f_{35}	f_{36}
`Country` No	✓		✓		✓		✓	✓			✓	✓			✓	✓		
`Country` Yes		✓		✓		✓			✓	✓			✓	✓			✓	✓
`Transformation` X	✓	✓					✓	✓	✓	✓								
`Transformation` $\log(X+1)$			✓	✓							✓	✓	✓	✓				
`Transformation` X^2					✓	✓									✓	✓	✓	✓
`Differences` Yes							✓		✓		✓		✓		✓		✓	
`Differences` No	✓	✓	✓	✓	✓	✓		✓		✓		✓		✓		✓		✓

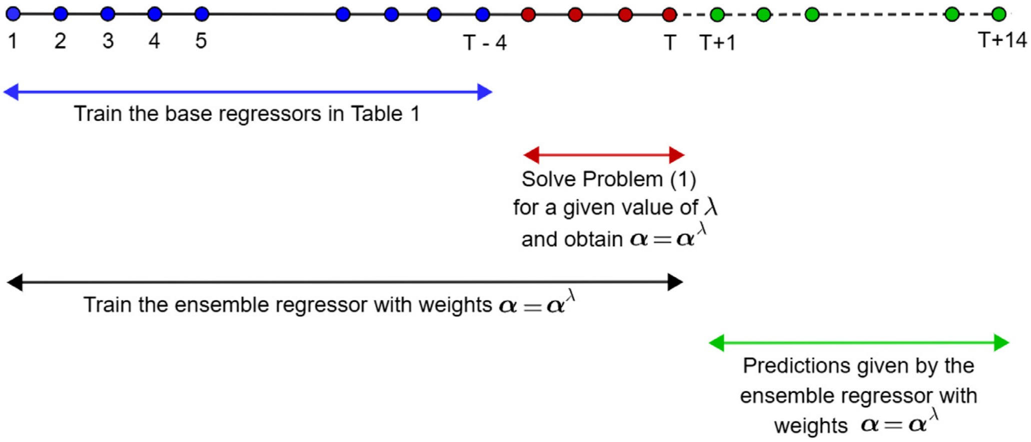

1
2
3
4
5
T - 4
T
T+1
T+14
Train the base regressors in Table 1
Solve Problem (1) for a given value of λ and obtain $\alpha = \alpha^{\lambda}$
Train the ensemble regressor with weights $\alpha = \alpha^{\lambda}$
Predictions given by the ensemble regressor with weights $\alpha = \alpha^{\lambda}$

Fig. 7. The timeline of building the base regressors in $\mathcal{F}$, solving Problem (1) to obtain the sparse ensemble for a given value of λ, and making the out-of-sample predictions.

```
Algorithm 1: Pseudocode for the complete procedure.
Input: $\{X_t,\ t = 1,\dots,T\}$, $\{X_t^r,\ t = 1,\dots,T\}$, $r = 2,\dots,R$, and $\mathcal{F}$ as in Table 1
Set $\mathcal{L}$ equal to the loss defined in (3)
Train the base regressors in $\mathcal{F}$ in $t \in \{1,\dots,T-4\}$
for $\lambda$ in $\{0, 2^{-10}, 2^{-9},\dots,2^{3}\}$ do
    Solve Problem (1) for $\lambda$ in $t \in \{T-3,\dots,T\}$ and obtain an optimal solution, $\alpha^{\lambda}$
end
Train the base regressors in $\mathcal{F}$ in $t \in \{1,\dots,T\}$
for $\lambda$ in $\{0, 2^{-10}, 2^{-9},\dots,2^{3}\}$ do
    Build the final ensemble regressor with weights $\alpha = \alpha^{\lambda}$
    Compute the predictions given by the final ensemble regressor with weights $\alpha = \alpha^{\lambda}$ in $t \in \{T+1,\dots,T+14\}$
end
Output: For each $\lambda$, the fourteen-days-ahead out-of-sample predictions of the final ensemble regressor with weights $\alpha = \alpha^{\lambda}$
```

the base regressors, namely, $T-3, T-2, T-1, T,$ while the individual losses are taken as $\mathcal{L}_f = \mathcal{L}(f)$. For each value of λ, we obtain the optimal weights α^{λ} returned by Problem (1). With these weights, the final ensemble regressor is built using all the data up to day T, and this final ensemble regressor is used to make fourteen-day-ahead predictions in $t \in \{T+1,\dots,T+14\}$.

The commercial optimization package Gurobi (Gurobi Optimization, 2018) has been used to solve the convex quadratic problems with linear constraints arising when solving Problem (1) with the loss in (3). Our experiments have been conducted on a PC, with an Intel ®Core$^{\text{TM}}$ i7-8550U CPU 1.80GHz processor and 8 GB RAM. The operating system is 64 bits.

3.5. The numerical results

The out-of-sample prediction performance of our approach is illustrated in three training and testing splits, with all training periods starting on 10/03/2020 for Andalusia and on 06/03/2020 for Sjælland, and all testing periods containing 14 days. For Andalusia, we have 10/03/2020–03/04/2020 (Training Period 1) and 04/04/2020–17/04/2020 (Testing Period 1), 10/03/2020–14/04/2020 (Training Period 2) and 15/04/2020–28/04/2020 (Testing Period 2), and 10/03/2020–06/05/2020 (Training Period 3) 07/05/2020–20/05/2020 (Testing Period 3). Similar periods are chosen for Sjælland, where all training periods start on 06/03/2020.

For each value of λ in the considered grid, the fourteen-days-ahead predictions made by the ensemble together with the realized values of the variable can be found in Tables 2–7 for each period and region, while Tables 8 and 9 report the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) over the fourteen days. In Tables 8 and 9, we highlight in bold the best MSE performance of the ensemble across all the values of λ considered, and denote by λ^{best} the value of the parameter where the minimum MSE is achieved. Note that in this case, for each period and region combination, the best MAE is also achieved at $\lambda = \lambda^{\text{best}}$. Figs. 14 and 15 present the weights of the base regressors in the ensembles as a function of λ by means of heatmaps. The color bar of each heatmap transitions from white to black, where the darker means a higher weight.

Figs. 8–13 depict the realized values of the variable at hand, cumulative number of hospitalized patients in the respective region (in red), as well as the fourteen-days-ahead predictions for three different ensembles. In the first ensemble, with $\lambda = 0$, the selective sparsity term does not play a role by construction (blue line). In the second ensemble, $\lambda = \lambda^{\text{best}}$, the ensemble is the one that performs the best in terms of MSE among all values of λ considered (black line). Finally, in the third ensemble, with $\lambda = \lambda^{\circ}$, the ensemble is the one showing the highest level of sparsity (green line).

We start by discussing the results obtained for Period 1 in Andalusia. In Fig. 8, we can see that it is possible to improve the out-of-sample prediction performance by taking a strictly positive value of λ. As pointed in the introduction, this is one of the advantages of our approach, namely, when seeking selective sparsity one may obtain also improvements on the out-of-sample prediction performance. A great benefit is observed with the ensemble that performs the best (black line), which is rather close to the actual values (red line). While the ensemble with $\lambda = 0$ presents a MAE of 532.71, for $\lambda^{\text{best}} = 2^{-6}$ the MAE is reduced to 40.50. This ensemble consists of the base regressors $f_2 \in \mathcal{F}_{\text{SVR}}$ and $f_{21}, f_{23} \in \mathcal{F}_{\text{LR}}$, with respective weights 0.71, 0.14 and 0.15. In Fig. 9, we plot the out-of-sample information for Andalusia and Period 2. Similar conclusions hold. In addition, the best ensemble is the one with $\lambda^{\text{best}} = 2^{-5}$, and consists of $f_5, f_{11} \in \mathcal{F}_{\text{SVR}}$, with respective weights 0.25 and 0.75. This means that the ensemble composition has changed over time, which can be explained by the nonstationarity of the data. If after having built the best ensemble for Training Period 1 one would have discarded these two base regressors because they were not selected, we would have lost the best combination for Training Period 2. This illustrates another advantage of our approach, namely, its adaptability. The ensemble composition changes again in Training Period 3 in Andalusia, where

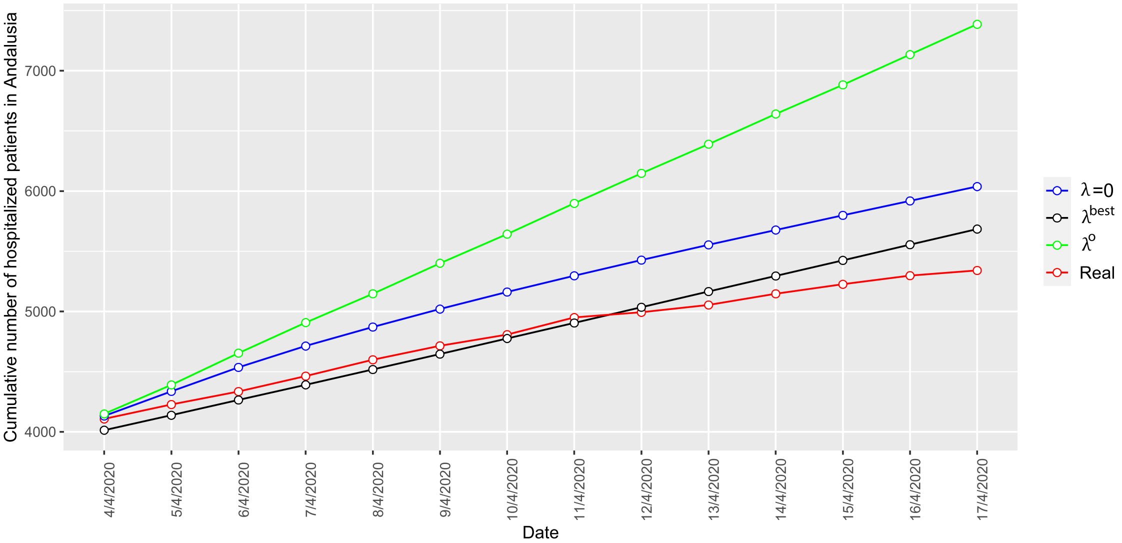


Fig. 8. Fourteen-day-ahead predictions for the cumulative number of hospitalized patients in Andalusia for COVID-19 in Testing Period 1 for three values of the tradeoff parameter λ, together with the actual values of the variable. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

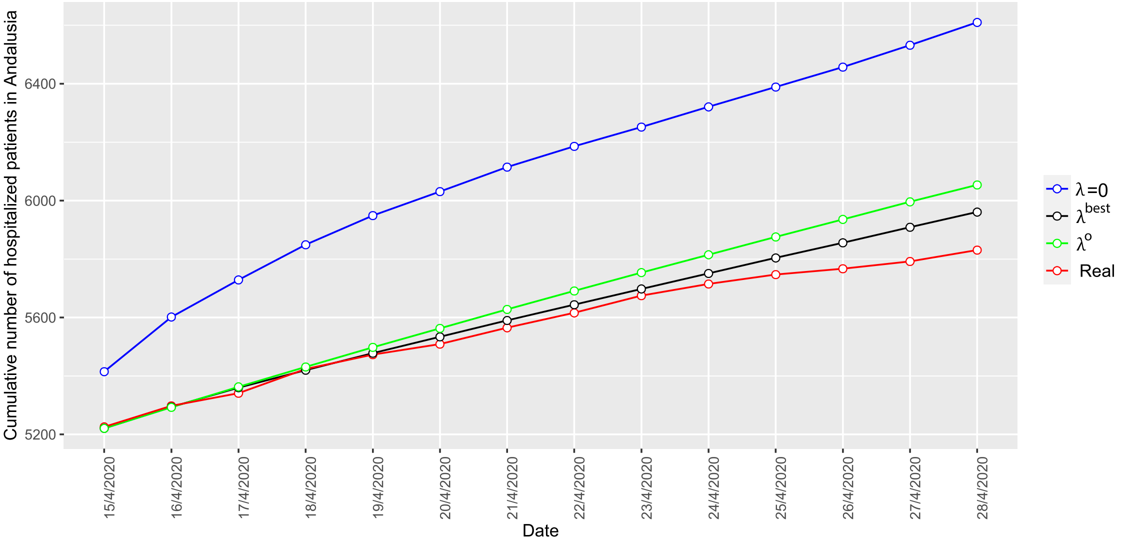


Fig. 9. Fourteen-day-ahead predictions for the cumulative number of hospitalized patients in Andalusia for COVID-19 in Testing Period 2 for three values of the tradeoff parameter λ, together with the actual values of the variable. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
For each value of λ, fourteen-day-ahead predictions of the ensemble for the cumulative number of hospitalized patients in Andalusia for COVID-19 in Testing Period 1. Last row shows the actual values.

λ	04/04	05/04	06/04	07/04	08/04	09/04	10/04	11/04	12/04	13/04	14/04	15/04	16/04	17/04
0	4132	4337	4536	4713	4871	5020	5162	5297	5427	5554	5677	5799	5919	6038
2^{-10}	4073	4233	4386	4527	4655	4776	4892	5005	5115	5225	5333	5442	5552	5662
2^{-9}	3985	4067	4146	4220	4290	4356	4419	4481	4541	4601	4659	4717	4775	4833
2^{-8}	3961	4021	4079	4132	4183	4231	4277	4321	4365	4407	4447	4488	4528	4568
2^{-7}	3960	4021	4078	4132	4183	4230	4277	4321	4364	4407	4447	4488	4527	4567
2^{-6}	3980	4064	4148	4228	4307	4385	4462	4537	4613	4688	4761	4835	4908	4981
2^{-5}	4014	4138	4265	4391	4518	4646	4776	4905	5035	5166	5295	5425	5555	5685
2^{-4}	4066	4246	4434	4628	4829	5037	5250	5468	5689	5911	6132	6351	6564	6772
2^{-3}	4121	4341	4579	4813	5040	5280	5514	5759	6001	6239	6482	6718	6957	7195
2^{-2}	4102	4302	4515	4722	4920	5127	5326	5534	5739	5938	6144	6343	6548	6754
2^{-1}	4106	4308	4524	4734	4935	5145	5348	5559	5767	5969	6178	6380	6588	6797
2^{0}	4112	4320	4543	4760	4966	5183	5391	5609	5822	6030	6245	6453	6668	6883
2^{1}	4125	4344	4581	4810	5028	5257	5477	5707	5934	6153	6381	6600	6827	7055
2^{2}	4149	4390	4654	4908	5147	5401	5643	5898	6148	6390	6641	6883	7134	7386
2^{3}	4149	4390	4654	4908	5147	5401	5643	5898	6148	6390	6641	6883	7134	7386
Actual	4107	4227	4335	4463	4599	4715	4808	4950	4993	5054	5147	5226	5298	5341

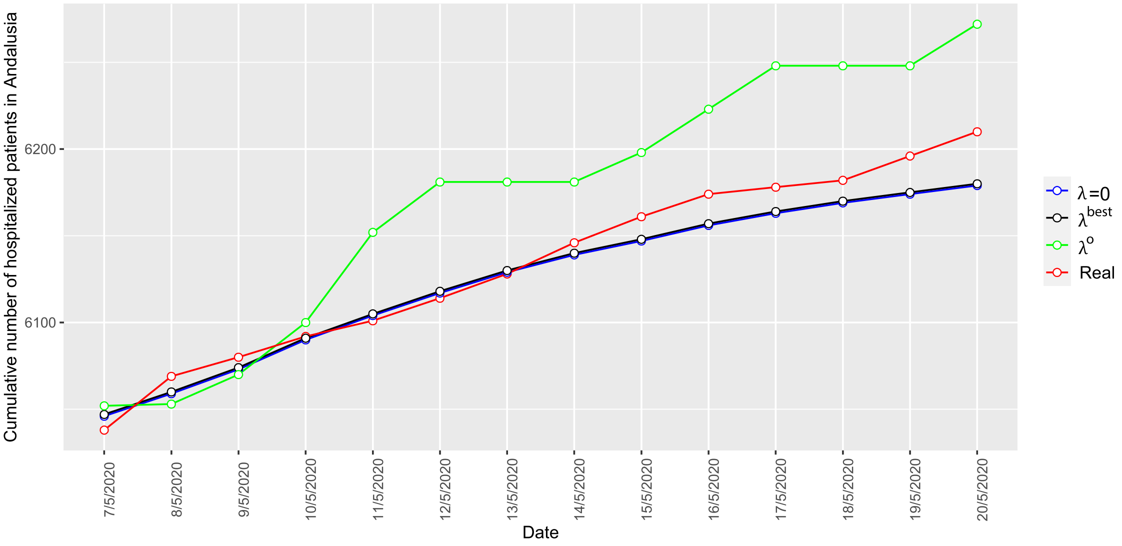


Fig. 10. Fourteen-day-ahead predictions for the cumulative number of hospitalized patients in Andalusia for COVID-19 in Testing Period 3 for three values of the tradeoff parameter λ, together with the actual values of the variable. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

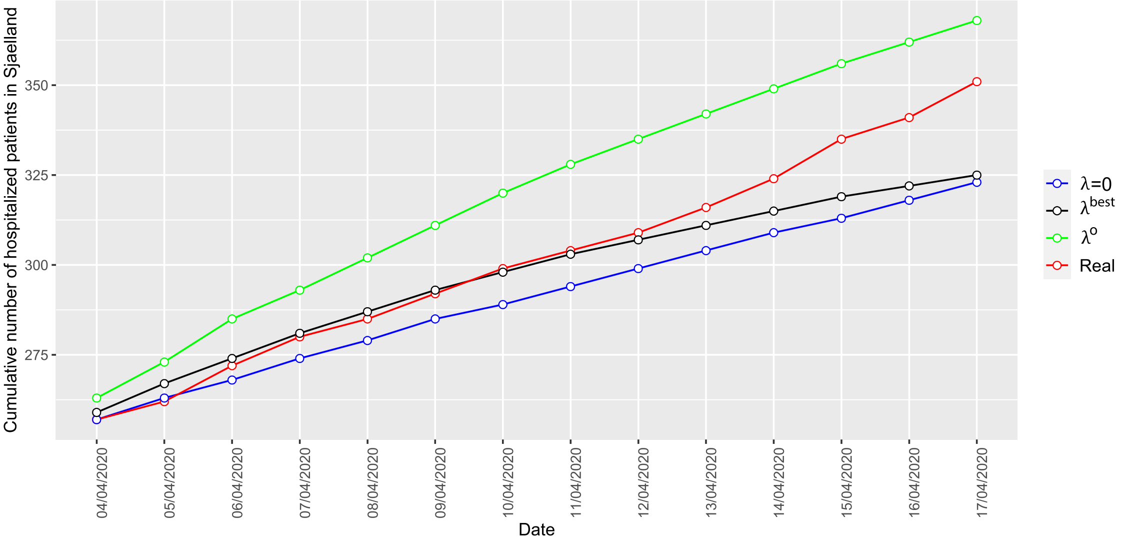


Fig. 11. Fourteen-day-ahead predictions for the cumulative number of hospitalized patients in Sjælland for COVID-19 in Testing Period 1 for three values of the tradeoff parameter λ, together with the actual values of the variable. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
For each value of λ, fourteen-day-ahead predictions of the ensemble for the cumulative number of hospitalized patients in Andalusia for COVID-19 in Testing Period 2. Last row shows the actual values.

λ	15/04	16/04	17/04	18/04	19/04	20/04	21/04	22/04	23/04	24/04	25/04	26/04	27/04	28/04
0	5415	5602	5729	5849	5949	6031	6115	6186	6252	6321	6389	6457	6532	6610
2^{-10}	5412	5598	5724	5843	5943	6025	6109	6179	6246	6315	6383	6451	6525	6603
2^{-9}	5411	5596	5722	5840	5939	6021	6105	6175	6241	6310	6378	6446	6520	6598
2^{-8}	5407	5590	5715	5832	5930	6011	6095	6165	6231	6300	6368	6436	6510	6587
2^{-7}	5346	5491	5597	5699	5787	5862	5939	6006	6070	6136	6201	6265	6334	6405
2^{-6}	5221	5294	5360	5420	5478	5534	5590	5644	5698	5751	5804	5856	5909	5961
2^{-5}	5219	5290	5359	5425	5490	5553	5616	5677	5738	5798	5857	5916	5975	6033
2^{-4}	5220	5290	5358	5424	5489	5551	5614	5675	5735	5794	5853	5911	5969	6024
2^{-3}	5220	5292	5361	5429	5495	5560	5624	5686	5749	5809	5870	5929	5989	6046
2^{-2}	5221	5293	5363	5431	5498	5563	5628	5691	5754	5815	5876	5936	5996	6054
2^{-1}	5221	5293	5363	5431	5498	5563	5628	5691	5754	5815	5876	5936	5996	6054
2^{0}	5221	5293	5363	5431	5498	5563	5628	5691	5754	5815	5876	5936	5996	6054
2^{1}	5221	5293	5363	5431	5498	5563	5628	5691	5754	5815	5876	5936	5996	6054
Actual	5226	5298	5341	5424	5473	5509	5565	5615	5675	5715	5747	5767	5792	5831

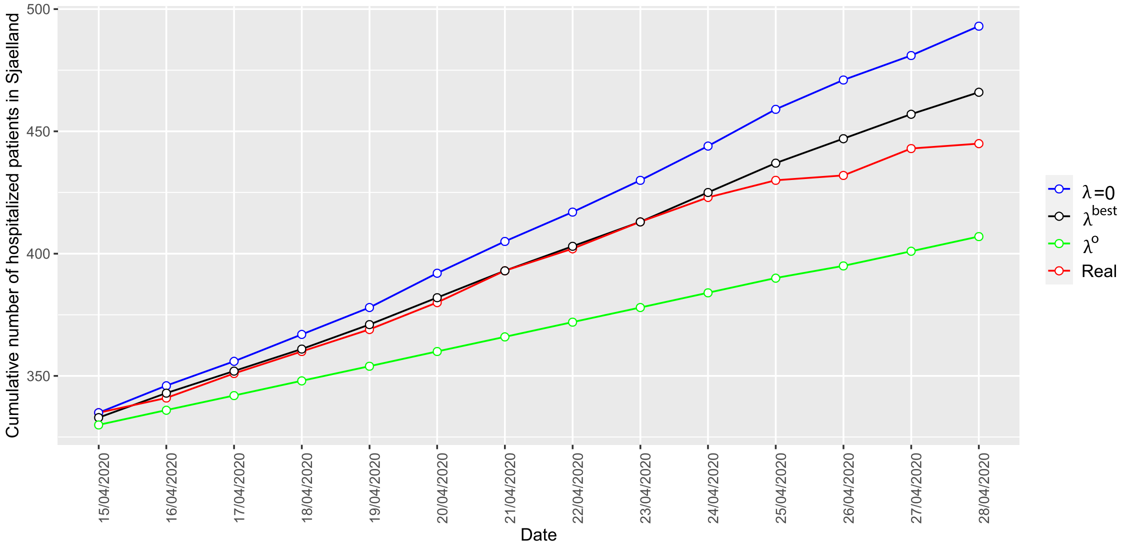


Fig. 12. Fourteen-day-ahead predictions for the cumulative number of hospitalized patients in Sjælland for COVID-19 in Testing Period 2 for three values of the tradeoff parameter λ, together with the actual values of the variable. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

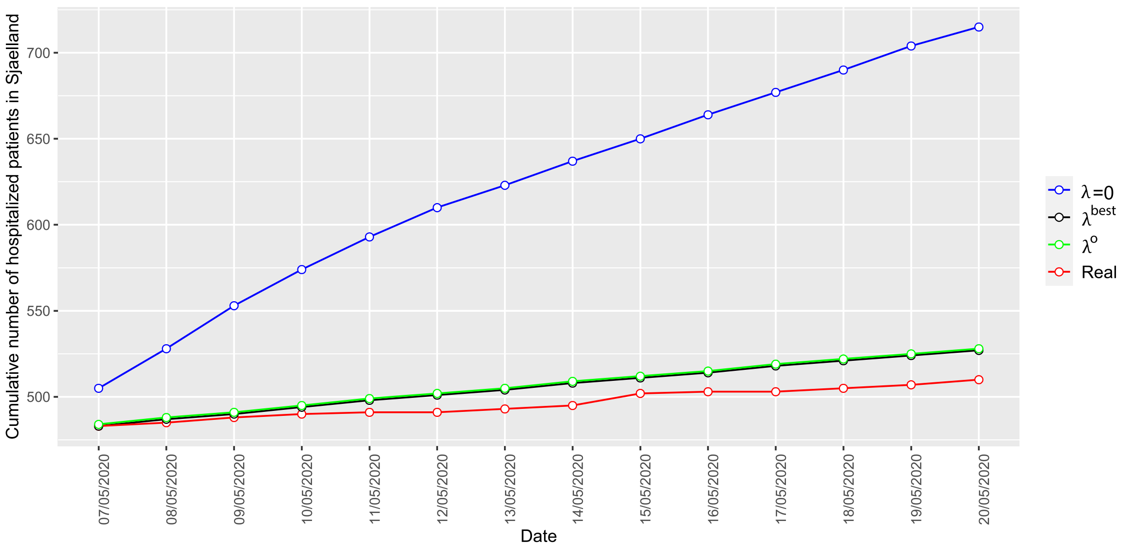


Fig. 13. Fourteen-day-ahead predictions for the cumulative number of hospitalized patients in Sjælland for COVID-19 in Testing Period 3 for three values of the tradeoff parameter λ, together with the actual values of the variable. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
For each value of λ, fourteen-day-ahead predictions of the ensemble for the cumulative number of hospitalized patients in Andalusia for COVID-19 in Testing Period 3. Last row shows the actual values.

λ	07/05	08/05	09/05	10/05	11/05	12/05	13/05	14/05	15/05	16/05	17/05	18/05	19/05	20/05
0	6046	6059	6073	6090	6104	6117	6129	6139	6147	6156	6163	6169	6174	6179
2^{-10}	6043	6054	6062	6067	6069	6069	6068	6063	6057	6047	6035	6020	6002	5982
2^{-9}	6043	6054	6062	6067	6069	6069	6068	6064	6057	6048	6035	6020	6003	5983
2^{-8}	6047	6063	6077	6089	6100	6110	6120	6128	6136	6142	6148	6152	6156	6159
2^{-7}	6039	6048	6055	6062	6069	6075	6082	6088	6094	6099	6104	6108	6112	6116
2^{-6}	6043	6054	6065	6074	6084	6092	6099	6106	6113	6120	6126	6131	6136	6141
2^{-5}	6045	6055	6066	6080	6098	6110	6116	6122	6131	6141	6151	6155	6159	6168
2^{-4}	6050	6056	6071	6091	6124	6144	6147	6151	6164	6180	6197	6199	6202	6218
2^{-3}	6049	6056	6070	6091	6125	6145	6148	6152	6165	6182	6199	6201	6204	6220
2^{-2}	6050	6056	6071	6093	6128	6148	6152	6156	6169	6187	6204	6207	6209	6226
2^{-1}	6050	6056	6071	6093	6129	6150	6153	6157	6170	6188	6206	6209	6211	6228
2^{0}	6050	6056	6071	6094	6131	6153	6156	6159	6173	6192	6210	6212	6214	6232
2^{1}	6051	6055	6071	6095	6135	6158	6161	6164	6178	6198	6218	6219	6221	6240
2^{2}	6051	6054	6070	6098	6144	6170	6171	6172	6188	6210	6233	6234	6235	6256
2^{3}	6052	6053	6070	6100	6152	6181	6181	6181	6198	6223	6248	6248	6248	6272
Actual	6038	6069	6080	6092	6101	6114	6128	6146	6161	6174	6178	6182	6196	6210

Table 5
For each value of λ, fourteen-day-ahead predictions of the ensemble for the cumulative number of hospitalized patients in Sjælland for COVID-19 in Testing Period 1. Last row shows the actual values.

λ	04/04	05/04	06/04	07/04	08/04	09/04	10/04	11/04	12/04	13/04	14/04	15/04	16/04	17/04
0	257	263	268	274	279	285	289	294	299	304	309	313	318	323
2^{-10}	259	266	272	279	285	290	295	299	304	307	311	314	317	319
2^{-9}	259	266	273	279	285	291	296	301	305	308	312	316	318	321
2^{-8}	259	267	274	281	287	293	298	303	307	311	315	319	322	325
2^{-7}	260	266	274	279	285	290	295	300	304	308	312	316	319	322
2^{-6}	261	271	281	289	297	305	313	321	327	334	340	347	353	358
2^{-5}	262	271	282	290	298	307	315	323	329	336	343	350	356	362
2^{-4}	262	272	283	291	300	309	318	326	333	340	347	354	360	367
2^{-3}	262	272	284	292	300	309	318	326	333	340	347	354	361	367
2^{-2}	262	272	284	292	301	310	319	327	334	341	348	355	361	367
2^{-1}	263	273	285	293	302	311	320	328	335	342	349	356	362	368
2^{0}	263	273	285	293	302	311	320	328	335	342	349	356	362	368
2^{1}	263	273	285	293	302	311	320	328	335	342	349	356	362	368
2^{2}	263	273	285	293	302	311	320	328	335	342	349	356	362	368
2^{3}	263	273	285	293	302	311	320	328	335	342	349	356	362	368
Actual	257	262	272	280	285	292	299	304	309	316	324	335	341	351

Table 6
For each value of λ, fourteen-day-ahead predictions of the ensemble for the cumulative number of hospitalized patients in Sjælland for COVID-19 in Testing Period 2. Last row shows the actual values.

λ	15/04	16/04	17/04	18/04	19/04	20/04	21/04	22/04	23/04	24/04	25/04	26/04	27/04	28/04
0	335	346	356	367	378	392	405	417	430	444	459	471	481	493
2^{-10}	334	345	355	365	376	389	401	413	424	438	452	463	473	484
2^{-9}	333	343	352	361	371	382	393	403	413	425	437	447	457	466
2^{-8}	331	339	346	353	361	369	376	384	391	400	408	416	424	431
2^{-7}	330	336	342	348	353	359	364	369	375	381	386	392	398	405
2^{-6}	330	336	342	347	353	359	365	370	376	382	388	393	399	405
2^{-5}	330	337	343	350	356	363	369	376	382	389	395	401	408	414
2^{-4}	330	337	343	349	356	362	369	375	382	388	395	400	407	414
2^{-3}	330	336	343	349	355	362	368	374	381	387	394	399	405	412
2^{-2}	330	336	342	348	355	361	367	373	379	385	391	396	403	409
2^{-1}	330	336	342	348	354	360	366	372	378	384	390	395	401	407
2^{0}	330	336	342	348	354	360	366	372	378	384	390	395	401	407
2^{1}	330	336	342	348	354	360	366	372	378	384	390	395	401	407
Actual	335	341	351	360	369	380	393	402	413	423	430	432	443	445

Table 7
For each value of λ, fourteen-day-ahead predictions of the ensemble for the cumulative number of hospitalized patients in Sjælland for COVID-19 in Testing Period 3. Last row shows the actual values.

λ	07/05	08/05	09/05	10/05	11/05	12/05	13/05	14/05	15/05	16/05	17/05	18/05	19/05	20/05
0	505	528	553	574	593	610	623	637	650	664	677	690	704	715
2^{-10}	504	526	550	571	589	607	619	633	646	659	673	685	698	710
2^{-9}	503	525	548	568	586	603	616	629	642	655	668	680	693	704
2^{-8}	501	522	544	563	580	596	608	621	634	646	659	671	683	694
2^{-7}	495	516	539	557	573	588	602	616	629	643	657	671	684	697
2^{-6}	491	510	535	550	565	579	592	605	617	631	645	659	672	686
2^{-5}	488	503	523	535	547	558	568	579	588	600	611	623	634	645
2^{-4}	483	490	500	505	511	517	522	527	533	539	545	551	557	563
2^{-3}	483	487	491	494	498	501	505	508	511	515	518	522	525	528
2^{-2}	483	487	491	494	498	501	505	508	511	515	518	521	524	527
2^{-1}	483	487	490	494	498	501	504	508	511	514	518	521	524	527
2^{0}	483	487	490	494	498	501	504	508	511	514	518	521	524	527
2^{1}	483	487	490	494	498	501	504	508	511	514	518	521	524	527
2^{2}	483	487	490	494	498	501	504	508	511	514	518	521	524	527
2^{3}	483	487	490	494	498	501	504	508	511	514	518	521	524	527
Actual	483	485	488	490	491	491	493	495	502	503	503	505	507	510

$f_{12} \in \mathcal{F}_{\mathrm{SVR}}$, $f_{15}, f_{16} \in \mathcal{F}_{\mathrm{RF}}$ and $f_{22} \in \mathcal{F}_{\mathrm{LR}}$ compose the best ensemble, see Fig. 10. Note that, for this particular period, λ^{best}=0, although this is not in general the case.

Regarding Sjælland, similar conclusions are obtained, see Table 9 and Figs. 11–13. The best ensembles are achieved for strictly positive values of λ, namely, $\lambda^{\mathrm{best}} = 2^{-9}$ for Testing Period 1, $\lambda^{\mathrm{best}} = 2^{-8}$ for Testing Period 2 and $\lambda^{\mathrm{best}} = 2^{-1}$ for Testing Period 3. Their compositions also differ across the three periods, $f_4, f_{11} \in \mathcal{F}_{\mathrm{SVR}}$, $f_{24} \in \mathcal{F}_{\mathrm{LR}}$ and $f_{26} \in \mathcal{F}_{\mathrm{S-ORRT}}$ for Training Period 1, $f_9 \in \mathcal{F}_{\mathrm{SVR}}$, $f_{23} \in \mathcal{F}_{\mathrm{LR}}$ and $f_{30}, f_{34} \in \mathcal{F}_{\mathrm{S-ORRT}}$ in Training Period 2, and $f_6, f_{10} \in \mathcal{F}_{\mathrm{SVR}}$ in Training Period 3. This again illustrates the advantage of our approach in terms of adaptability.

We end the section with a few words about the set of base regressors. In their last row, Tables 8 and 9 report the MSE and MAE of a persistence model in which the increase in the variable is kept constant throughout the testing period and equal to the last increase in the training period. As for any forecasting model, the persistence model might yield good results in some cases, such as

Table 8
For each value of λ, Mean Squared Error (MSE) and Mean Absolute Error (MAE) of the ensemble for Testing Period 1, 2 and 3 in Andalusia. For each period, the best performance is highlighted in bold. Last row contains the MSE and MAE of the persistence model tested.

	Testing Period 1		Testing Period 2		Testing Period 3	
	(04/04/2020–17/04/2020)		(15/04/2020–28/04/2020)		(07/05/2020–20/05/2020)	
λ	MSE	MAE	MSE	MAE	MSE	MAE
0	309188.29	532.71	174813.93	372.79	**188.86**	**11.00**
2^{-10}	302755.21	526.93	22697.21	120.07	12713.93	88.64
2^{-9}	298320.21	523.07	154944.93	369.50	12623.93	88.36
2^{-8}	288510.14	514.14	311559.21	518.21	585.00	18.57
2^{-7}	151329.79	368.50	311996.36	518.64	3353.57	51.43
2^{-6}	**3290.07**	**40.50**	105662.43	311.86	1635.64	35.36
2^{-5}	9174.07	71.21	**21515.93**	**118.07**	565.00	20.43
2^{-4}	8477.29	68.29	554612.29	585.43	214.64	12.21
2^{-3}	10905.86	78.86	1034287.57	841.14	243.57	13.29
2^{-2}	11893.29	82.86	580431.07	625.79	351.07	16.50
2^{-1}	11893.29	82.86	620786.79	648.36	397.14	17.57
2^{0}	11893.29	82.86	705260.29	694.43	498.00	19.86
2^{1}	11893.29	82.86	890921.71	786.86	737.07	24.36
2^{2}	11893.29	82.86	1310319.64	964.93	1387.93	33.36
2^{3}	11893.29	82.86	1310319.64	964.93	2236.71	42.14
Persistence	3243399.00	1429.89	183250.60	347.38	228.52	10.98

Table 9
For each value of λ, Mean Squared Error (MSE) and Mean Absolute Error (MAE) of the ensemble for Testing Period 1, 2 and 3 in Sjælland. For each period, the best performance is highlighted in bold. Last row contains the MSE and MAE of the persistence model tested.

	Testing Period 1		Testing Period 2		Testing Period 3	
	(04/04/2020–17/04/2020)		(15/04/2020–28/04/2020)		(07/05/2020–20/05/2020)	
λ	MSE	MAE	MSE	MAE	MSE	MAE
0	538.07	18.36	186.00	11.00	19171.64	126.93
2^{-10}	327.50	14.07	170.14	8.71	18097.14	123.14
2^{-9}	**66.71**	**5.00**	146.79	7.93	17080.29	119.57
2^{-8}	228.71	13.43	**103.14**	**6.57**	15200.14	112.57
2^{-7}	947.93	27.07	141.79	8.21	14582.07	108.64
2^{-6}	905.43	26.57	164.14	12.14	12379.21	99.36
2^{-5}	600.29	21.71	217.79	14.07	7189.43	75.43
2^{-4}	622.57	22.14	313.21	16.79	1055.79	28.36
2^{-3}	671.57	23.00	319.29	17.00	134.14	10.00
2^{-2}	761.57	24.43	343.14	17.57	126.79	9.79
2^{-1}	818.00	25.29	379.29	18.57	**123.14**	**9.57**
2^{0}	818.00	25.29	379.29	18.57	123.14	9.57
2^{1}	818.00	25.29	379.29	18.57	123.14	9.57
2^{2}	818.00	25.29	379.29	18.57	123.14	9.57
2^{3}	818.00	25.29	379.29	18.57	123.14	9.57
Persistence	593.83	19.88	36.38	4.92	5.15	1.91

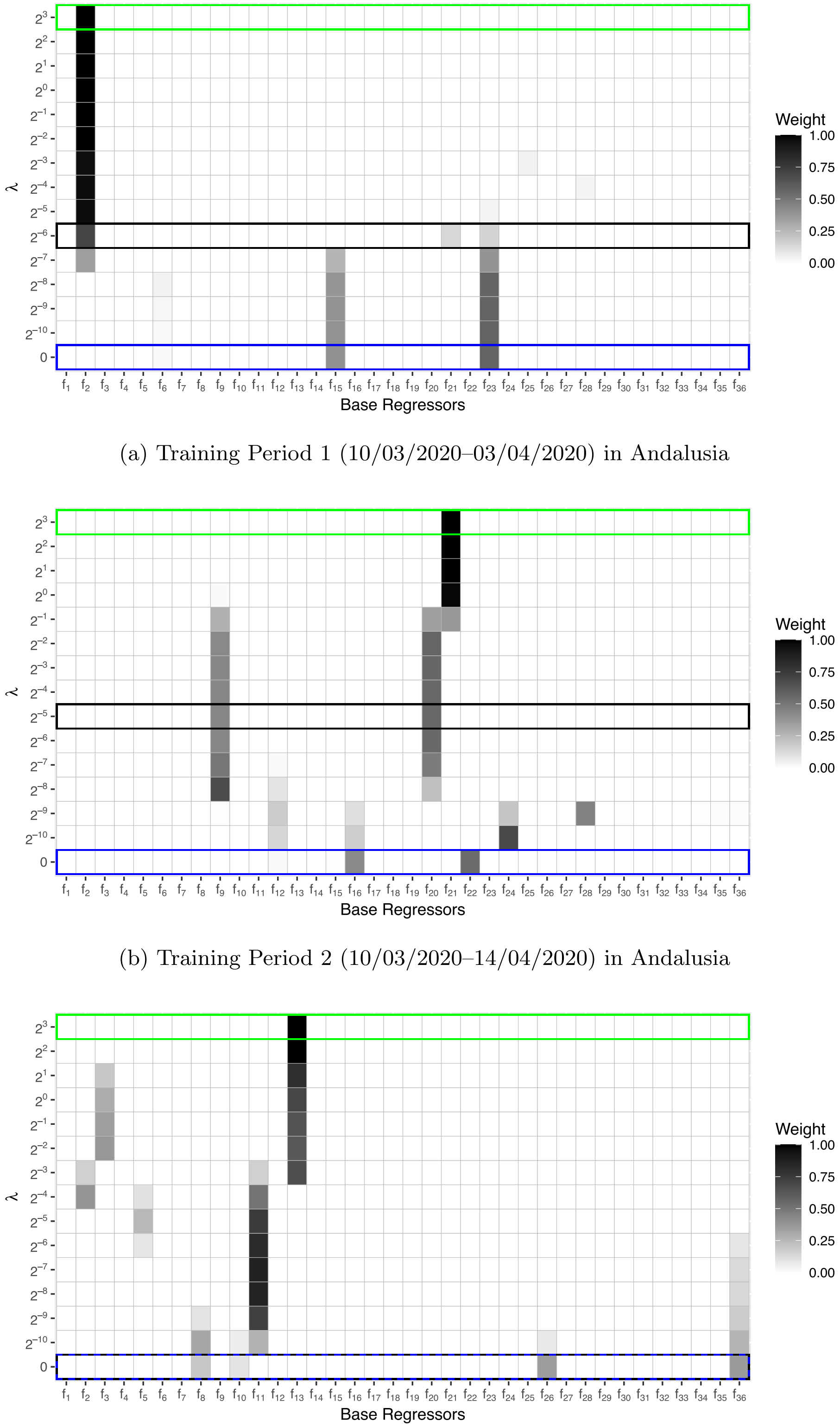
in Testing Period 2 and 3 in Sjælland, but very poor ones in other situations, such as in Testing Period 1 and 2 in Andalusia. We could have easily embedded this persistence model, or any other one, by enlarging the set of base regressors. Again, because of the adaptability of our approach, the persistence model would have been chosen or not to be part of the sparse ensemble, depending on the period and the region being considered.

4. Conclusions

In this paper we have addressed the problem of building ensembles with selective sparsity of regression methods, which is suitable in changing circumstances such as those related to the COVID-19 pandemic. The construction of the ensemble amounts to solving an optimization problem, which is quadratic convex under linear constraints for the empirical Ordinary Least Squares regression loss and it can be written as a linear problem for empirical loss of quantile regression. Under convexity assumptions on the loss $\mathcal{L}$, we show that, by varying the parameter λ in the interval $[0, \lambda^{\circ}]$ we move from the ensemble minimizing the overall loss $\mathcal{L}$ to the ensemble with one single base regressor f, namely, the one with lowest individual loss $\mathcal{L}_f$. Moreover, different types of desirable properties of the ensemble can be easily accommodated by modifying the penalty term or the constraints. The application to data on hospitalized patients in Andalusia (Spain) and Sjælland (Denmark) shows the advantage of using an ensemble with selective sparsity instead of a rough ensemble or one single base regressor.

The computational experience reported is limited to the problem motivating this work. For other types of problems, it may be interesting to combine the selective sparsity suggested in this paper (number of regressors used) with the feature sparsity (number of features used), by adding ℓ_∞ penalties as in Section 2.3 and in Blanquero et al. (2020a) and Blanquero, Carrizosa, Molero-Río, and Romero Morales (2020b). It may also be attractive to use different measures for the individual losses $\mathcal{L}_f$ and the overall loss $\mathcal{L}$. For instance, one can build the ensemble with lowest least squares errors, but being reluctant to use base regressors with high least absolute deviations, or more generally, quantile errors.

Even if we knew the probabilistic mechanism generating the data, sound probability assessments are rather difficult in the setting considered in this paper. Those probability assessments are what Efron (2020) calls “attributions”. As recognized in that paper, prediction is much easier than attribution. The use of an ad-

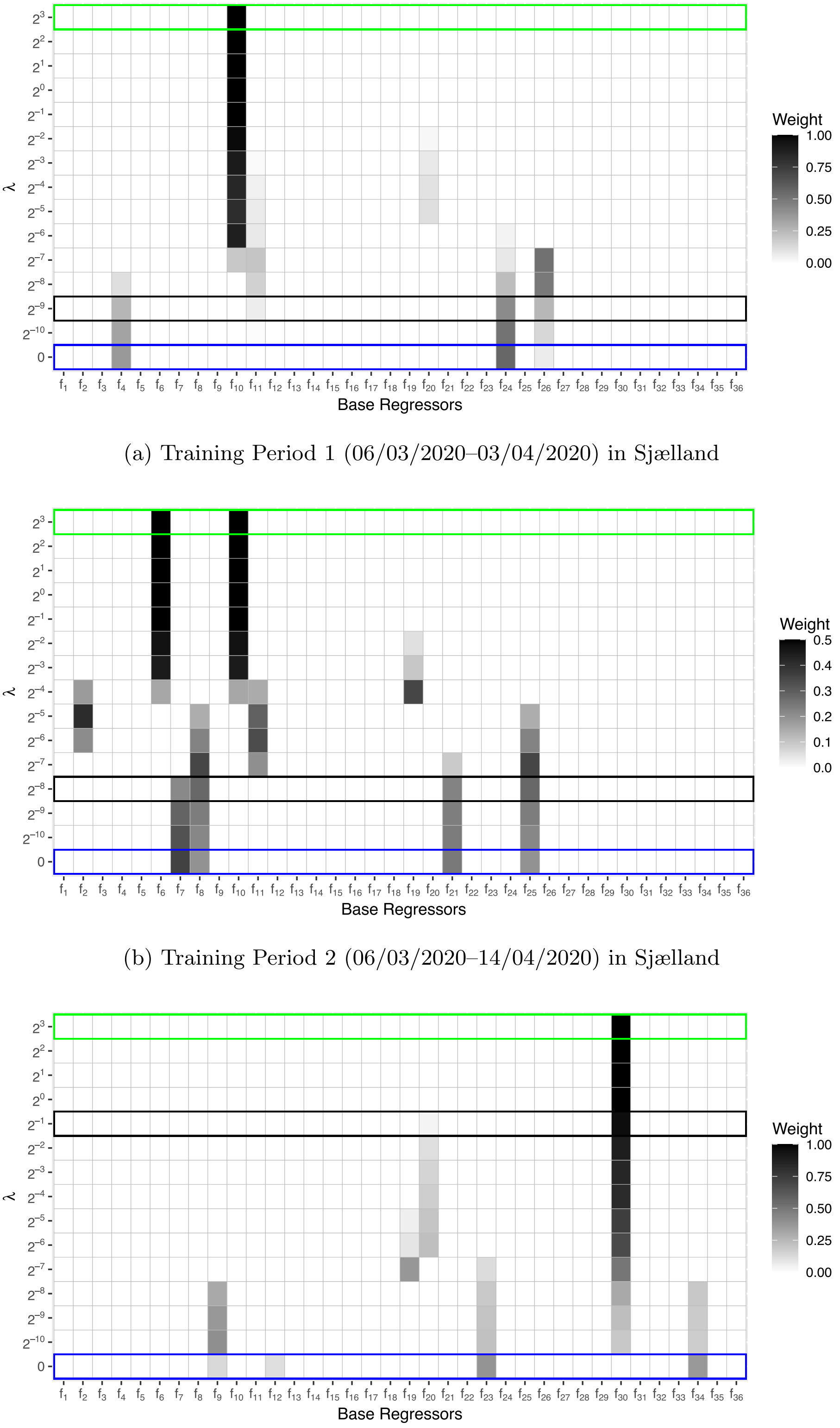


(a) Training Period 1 (10/03/2020–03/04/2020) in Andalusia

(b) Training Period 2 (10/03/2020–14/04/2020) in Andalusia

(c) Training Period 3 (10/03/2020–06/05/2020) in Andalusia

Fig. 14. For each value of λ, heatmap of the weights of the base regressors in the ensemble in Training Period 1, 2 and 3 in Andalusia. We highlight $\lambda = 0$ in blue, λ^{best} in black, and $\lambda = \lambda^{\circ}$ in green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) Training Period 1 (06/03/2020–03/04/2020) in Sjælland

(b) Training Period 2 (06/03/2020–14/04/2020) in Sjælland

(c) Training Period 3 (06/03/2020–06/05/2020) in Sjælland

Fig. 15. For each value of λ, heatmap of the weights of the base regressors in the ensemble in Training Period 1, 2 and 3 in Sjælland. We highlight $\lambda = 0$ in blue, λ^{best} in black, and $\lambda = \lambda^{\circ}$ in green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

equate bootstrap procedure (see Bühlmann, 2002, for a review of bootstraps for time series) could yield probability attributes. The consistency of the bootstrap for Support Vector Machines when the data can be assumed to be independent and identically distributed, has been shown in Christmann and Hable (2013). To the best of our knowledge, an analogous result for time series in a general setting as the one considered here has not been stated yet, and it certainly constitutes a field for future research.

Another challenging line of research is the construction of sparse ensembles (sparse both in base regressors and in features) for classification problems. Although some attempts have been made to address this problem using Linear Programming, Zhang and Zhou (2011), natural losses yield versions of Problem (1) with (many) binary variables, and thus new strategies are to be defined to cope with data sets of realistic size. This challenging problem is now under study.

Acknowledgements

We thank the reviewers for their thorough comments and suggestions, which have been very valuable to strengthen the quality of the paper. This research has been financed in part by research projects EC H2020 MSCA RISE NeEDS (Grant agreement ID: 822214); FQM-329 and P18-FR-2369 (Junta de Andalucía, Spain); MTM2017-89422-P (Ministerio de Economía, Industria y Competitividad, Spain); PID2019-110886RB-I00 (Ministerio de Ciencia, Innovacin y Universidades, Spain); PR2019-029 (Universidad de Cádiz, Spain); PITUFLOW-CM-UC3M (Comunidad de Madrid and Universidad Carlos III de Madrid, Spain); and EP/R00370X/1 (EPSRC, United Kingdom). This support is gratefully acknowledged.

References

Achterberg, M., Prasse, B., Ma, L., Trajanovski, S., Kitsak, M., & Van Mieghem, P. (2020). Comparing the accuracy of several network-based COVID-19 prediction algorithms. Forthcoming in International Journal of Forecasting.

Ando, T., & Li, K. C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association, 109*, 254–265.

Bates, J., & Granger, C. (1969). The combination of forecasts. *Operations Research Quarterly, 20*, 451–468.

Benítez-Peña, S., Blanquero, R., Carrizosa, E., & Ramírez-Cobo, P. (2019a). Cost-sensitive feature selection for support vector machines. *Computers & Operations Research, 106*, 169–178.

Benítez-Peña, S., Blanquero, R., Carrizosa, E., & Ramírez-Cobo, P. (2019b). On support vector machines under a multiple-cost scenario. *Advances in Data Analysis and Classification, 13*, 663–682.

Benítez-Peña, S., Blanquero, R., Carrizosa, E., & Ramírez-Cobo, P. (2020a). Cost-sensitive probabilistic predictions for support vector machines. *Technical Report IMUS, Sevilla, Spain*. https://www.researchgate.net/publication/341103637_Cost-sensitive_probabilistic_predictions_for_support_vector_machines.

Benítez-Peña, S., Carrizosa, E., Guerrero, V., Jiménez-Gamero, M. D., Martín-Barragán, B., Molero-Río, C., Ramírez-Cobo, P., Romero Morales, D., & Sillero-Denamiel, M. R. (2020b). Short-term predictions of the evolution of COVID-19 in andalusia. an ensemble method. *Technical Report IMUS, Sevilla, Spain*. https://www.researchgate.net/publication/340716304_Short-Term_Predictions_of_the_Evolution_of_COVID-19_in_Andalusia_An_Ensemble_Method.

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning, 106*, 1039–1082.

Bertsimas, D., King, A., & Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics, 44*, 813–852.

Blanquero, R., Carrizosa, E., Molero-Río, C., & Romero Morales, D. (2020a). On sparse optimal regression trees. *Technical Report IMUS, Sevilla, Spain*. https://www.researchgate.net/publication/341099512_On_Sparse_Optimal_Regression_Trees.

Blanquero, R., Carrizosa, E., Molero-Río, C., & Romero Morales, D. (2020b). Sparsity in optimal randomized classification trees. *European Journal of Operational Research, 284*, 255–272.

Blanquero, R., Carrizosa, E., Molero-Río, C., & Romero Morales, D. (2021). Optimal randomized classification trees. *Computers & Operations Research, 132*, 105281.

Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., & Sillero-Denamiel, M. R. (2020). A cost-sensitive constrained lasso. *Advances in Data Analysis and Classification, 15*, 121–158.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Bühlmann, P. (2002). Bootstraps for time series. *Statistical Science, 17*, 52–72.

Carrizosa, E., Martín-Barragán, B., & Romero Morales, D. (2008). Multi-group support vector machines with measurement costs: A biobjective approach. *Discrete Applied Mathematics, 156*, 950–966.

Carrizosa, E., Martín-Barragán, B., & Romero Morales, D. (2010). Binarized support vector machines. *INFORMS Journal on Computing, 22*, 154–167.

Carrizosa, E., Martín-Barragán, B., & Romero Morales, D. (2011). Detecting relevant variables and interactions in supervised classification. *European Journal of Operational Research, 213*, 260–269.

Carrizosa, E., Molero-Río, C., & Romero Morales, D. (2021). Mathematical optimization in classification and regression trees. *TOP, 29*, 5–33.

Carrizosa, E., Mortensen, L. H., Romero Morales, D., & Sillero-Denamiel, M. R. (2020a). On linear regression models with hierarchical categorical variables. *Technical Report IMUS, Sevilla, Spain*. https://www.researchgate.net/publication/341042405_On_linear_regression_models_with_hierarchical_categorical_variables.

Carrizosa, E., Nogales-Gómez, A., & Romero Morales, D. (2016). Strongly agree or strongly disagree?: Rating features in support vector machines. *Information Sciences, 329*, 256–273.

Carrizosa, E., Nogales-Gómez, A., & Romero Morales, D. (2017a). Clustering categories in support vector machines. *Omega, 66*, 28–37.

Carrizosa, E., Olivares-Nadal, A., & Ramírez-Cobo, P. (2017b). A sparsity-controlled vector autoregressive model. *Biostatistics, 18*, 244–259.

Carrizosa, E., Olivares-Nadal, A., & Ramírez-Cobo, P. (2020b). Novel constraints for enhancing interpretability in linear regression. *SORT (Statistics and Operations Research Transactions), 44*, 67–98.

Carrizosa, E., & Romero Morales, D. (2013). Supervised classification and mathematical optimization. *Computers and Operations Research, 40*, 150–165.

Christmann, A., & Hable, R. (2013). On the consistency of the bootstrap approach for support vector machines and related kernel-based methods. In B. Schölkopf, Z. Luo, & V. Vovk (Eds.), *Empirical inference: Festschrift in honor of vladimir n. vapnik* (pp. 231–244). Berlin, Heidelberg: Springer.

Datta, S., & Das, S. (2015). Near-Bayesian support vector machines for imbalanced data classification with equal or unequal misclassification costs. *Neural Networks, 70*, 39–52.

Deng, H. (2019). Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics, 7*, 277–287.

Efron, B. (2020). Prediction, estimation, and attribution. *Journal of the American Statistical Association, 115*, 636–655.

Fernández-Casal, R. (2020). COVID-19 github repository. Accessed on: September. https://github.com/rubenfcasal/COVID-19.

Florez-Lopez, R., & Ramon-Jeronimo, J. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. a correlated-adjusted decision forest proposal. *Expert Systems with Applications, 42*, 5737–5753.

Fountoulakis, K., & Gondzio, J. (2016). A second-order method for strongly convex ℓ_1-regularization problems. *Mathematical Programming, 156*, 189–219.

Friese, M., Bartz-Beielstein, T., Bäck, T., Naujoks, B., & Emmerich, M. (2019). Weighted ensembles in model-based global optimization. In *AIP conference proceedings*.

Friese, M., Bartz-Beielstein, T., & Emmerich, M. (2016). Building ensembles of surrogate models by optimal convex combination. *Technical Report*. http://nbn-resolving.de/urn:nbn:de:hbz:832-cos4-3480.

Gaines, B. R., Kim, J., & Zhou, H. (2018). Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics, 27*, 861–871.

Gambella, C., Ghaddar, B., & Naoum-Sawaya, J. (2021). Optimization models for machine learning: A survey. *European Journal of Operational Research, 290*, 807–828.

Gurobi Optimization, L. (2018). Gurobi optimizer reference. *manual*. http://www.gurobi.com.

Härdle, W. (1990). *Applied nonparametric regression, 19*. Cambridge University Press.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. CRC press.

Statens Serum Institut. (2020). COVID-19 SSI repository. Accessed on: September. https://covid19.ssi.dk/overvagningsdata.

Kedem, B., & Fokianos, K. (2005). *Regression models for time series analysis volume 488*. John Wiley & Sons.

Koenker, R., & Hallock, K. (2001). Quantile regression. *Journal of Economic Perspectives, 15*, 143–156.

Koenker, R., & Ng, P. (2005). Inequality constrained quantile regression. *Sankhyā: The Indian Journal of Statistics, 67*, 418–440.

Lee, Y., Nelder, J., & Pawitan, Y. (2018). Generalized linear models with random effects: Unified analysis via H-likelihood. *CRC Press, 153*.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News, 2*, 18–22.

Martín-Barragán, B., Lillo, R., & Romo, J. (2014). Interpretable support vector machines for functional data. *European Journal of Operational Research, 232*, 146–155.

Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. D. (2012). Ensemble approaches for regression: A survey. *ACM Computing Surveys, 45*, 1–40.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2019). e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071). *TU Wien*. https://CRAN.R-project.org/package=e1071R package version 1.7-1

Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., & Vasilakis, C. (2021). Forecasting and planning during a pandemic: Covid-19 growth rates, supply chain disruptions, and governmental decisions. *European Journal of Operational Research, 290*, 99–115.

Ren, Y., Zhang, L., & Suganthan, P. (2016). Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational Intelligence Magazine, 11*, 41–53.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag.

Zhang, L., & Zhou, W. D. (2011). Sparse ensembles using weighted combination methods based on linear programming. *Pattern Recognition, 44*, 97–106.