Contents lists available at ScienceDirect

International Journal of Forecasting

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

Hierarchical transfer learning with applications to electricity load forecasting

Anestis Antoniadis [a], Solenne Gaucher [b], Yannig Goude [c,d,*]

[a] Istituto di Scienze Applicate e Sistemi Intelligenti, Consiglio Nazionale delle Ricerche, Via Pietro Castellino 111, Napoli 80131 , Italy
[b] INRIA, Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, Bâtiment 307, 91405 Orsay, France
[c] EDF R&D, Saclay, 7 bd Gaspard Monge 91120 Palaiseau, France
[d] Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, Bâtiment 307, 91405 Orsay, France

ARTICLE INFO

Keywords:
Demand forecasting
Semi-parametric additive model
Random forest
Transfer learning
Time series
Combining forecasts
Aggregation of experts

ABSTRACT

The recent abundance of electricity consumption data available at different scales provides new opportunities and highlights the need for new techniques to leverage information present at finer scales in order to improve forecasts at wider scales. In this study, we take advantage of the similarity between this hierarchical prediction problem and transfer learning where source data are observed at a low aggregation level and target data at a global level. We develop two methods for hierarchical transfer learning based on stacking generalized additive models and random forests (GAM-RF). We also propose and compare adaptations of online aggregation of experts in a hierarchical context using quantile GAM-RF as experts. We apply these methods to two electricity load forecasting problems at the national scale by using smart meter data in the first case and regional data in the second case. For these two user cases, we compared the performance of our methods and benchmark algorithms, and investigated their behavior using variable importance analysis. Our results demonstrate that both methods can lead to significantly improved predictions.

1. Introduction

The recent abundance of electricity consumption data available at a low aggregation level due partly to the development of smart meters provides many new opportunities for electricity consumption forecasting (e.g., see Wang, Chen, Hong, and Kang 2019). However, these new perspectives come with new challenges, such as how to use these data obtained at a finer scale (corresponding to a household or a smaller geographical area), which can be used to create forecasts at a fine scale, into making predictions at a wider scale (e.g., at the national scale). In this study, we propose two methods for leveraging our ability to predict a variable of interest at a finer scale, with the goal of exploiting these predictions to improve prediction at a larger scale. This problem involves taking advantage of the similarities (e.g., similar dependency to explanatory variables and common drift in distributions) between forecasting problems at different scales and it can be naturally formulated in the framework of *transfer learning*.

Transfer learning methods aim to transfer the knowledge acquired from solving given problems (referred to as *source* problems $\mathcal{S}$) to address another problem of interest (referred to as the *target* problem $\mathcal{T}$). In a supervised predictive machine learning setting, the objective is to predict a variable of interest $Y^{\mathcal{T}}$ using covariates $\mathbf{X}^{\mathcal{T}}$. In particular, the learner employs a set of observations, $(\mathbf{X}_i^{\mathcal{T}}, Y_i^{\mathcal{T}})_{i\in\{1,\dots,n^{\mathcal{T}}\}}$, drawn from a joint distribution, $\mathbb{P}^{\mathcal{T}}$. Popular methods involve minimizing the empirical risk

[*] Corresponding author at: Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, Bâtiment 307, 91405 Orsay, France.

E-mail addresses: antoniadis.anestis@isasi.cnr.it (A. Antoniadis), solenne.gaucher@math.u-psud.fr (S. Gaucher), yannig.Goude@edf.fr (Y. Goude).

https://doi.org/10.1016/j.ijforecast.2023.04.006

corresponding to a given loss function over a set of possible learners (e.g., tree based, neural nets, and generalized additive models (GAMs)). If the learners and loss are selected in an appropriate manner, the estimated model will have good forecasting accuracy with a new data set provided that the size of the training set is sufficiently large, and the marginal and joint distributions remain unchanged in the test set. However, these conditions are not satisfied in a wide range of real-world applications. Classical examples of these situations include tasks that require massive training sets, such as computer vision or natural language processing. When dealing with temporal data, we may be confronted with changes in the distribution that lead to large prediction errors over the prediction period. If we consider that the historical data correspond to a first problem and the prediction period to a different but related problem, we must address the following challenge: exploiting the similarity between the two tasks in order to take advantage of the abundance of historical data while ensuring adaptability to the new task.

Transfer learning aims to tackle this problem and it has attracted increasing attention in machine learning, with many applications (see Olivas, Guerrero, Sober, Benedito, and Lopez 2009). In many practical situations, a relatively small quantity of data is available from the target distribution $\mathbb{P}^{\mathcal{T}}$. In some cases, we also have access to a larger data set with a different distribution denoted by $\mathbb{P}^{\mathcal{S}}$, which can be used to solve a task related to the target problem. A key assumption is that $\mathbb{P}^{\mathcal{T}}$ and $\mathbb{P}^{\mathcal{S}}$ are related in a way that can be leveraged by the transfer learning method. These distributions may be defined on the same domain (the transfer learning problem is said to be homogeneous in this case) or on different domains, and thus more complex transformations need to be developed (the transfer learning problem is said to be heterogeneous in this case). In this study, we focus on heterogeneous transfer problems with differences between the source and target in terms of the feature spaces, feature marginal distribution, and joint distribution.

Surprisingly, although transfer learning is very popular in computer vision and text mining (for a survey, see Pan and Yang 2010 and Zhuang et al. 2020), very few applications have been developed in the time series forecasting community. For example, Laptev, Yu, and Rajagopal (2018) fine-tuned a pre-trained neural network using a large data set of individual electricity loads as the source and some independent individual data as the target. In addition, Capezza, Palumbo, Goude, Wood, and Fasiolo (2021) proposed additive stacking (an interpretable aggregation of experts) by combining models at the individual and global levels for the probabilistic forecasting of individual demands. Moreover, Obst, Ghattas, Claudel, Cugliari, Goude, and Oppenheim (2022), Obst, de Vilmarest and Goude (2021) proposed a fine-tuning approach and used online updates to transfer information from Italian data to French data in order to improve electricity load forecasts during the COVID-19 lockdown. In this case, transfers occurred both in time (from past data (source) to future data (target)) and space (from one country to another).

Hybridizing statistical models with modern machine learning tools has recently been shown to be an efficient strategy for forecasting time series (see Smyl 2020 for the top rank in the M4 competition). Anderer and Li (2022) proposed transferring time-series features for bottom-up forecasting in the context of the M5 competition with intermittent time series at the low aggregation level. N-BEATS was proposed as a deep learning forecasting approach at the global level and by boosting trees with LightGBM at the bottom level. In the present study, we propose new transfer learning methods for hierarchical prediction by leveraging data available at a fine scale to improve prediction at a wider scale. The first approach (presented in detail in Section 3.1) is based on the design of new features learned from the source data. These features are then used as inputs in a random forest (RF) stacked with a GAM. Stacking an ensemble of forecasting models for time series forecasting has already been conducted successfully (e.g., see Khairalla, Ning, Al-Jallad, and El-Faroug 2018, Moon, Jung, Rew, Rho, and Hwang 2020 for load forecasting, (Dong, Zhang, Wang, & Zhou, 2021) wind power forecasting, (Xenochristou & Kapelan, 2020) water demand forecasting and (Zhai & Chen, 2018) pollution forecasting). Previous studies usually combined the forecasts in a meta-learner, whereas we propose combining features from GAM with the original covariates in the RF. This new method allows the detection of interactions that are not modeled in GAMs or that appear online (typically interaction with time in the context of drift). Our empirical results suggest that the proposed feature design method by combining stacked GAMs and RFs can improve predictions at a wider scale by leveraging knowledge acquired from data at a finer scale. Unfortunately, this approach relies on knowledge acquired from a training set, which may not be relevant if the distribution changes during the test period. Thus, these methods are not adaptive to abrupt changes in distribution both at fine and wider scales, and they cannot leverage the relationship between both. Therefore, to ensure the adaptivity of our model, we propose a second transfer learning approach based on the online aggregation of experts. In the hierarchical context, Brégère and Huard (2022) and Goehry, Goude, Massart, and Poggi (2019) showed that aggregating experts designed on different nodes of a hierarchical partition of the data (statistical clustering based on temporal or exogenous information or spatial partition) can improve the forecasting performance compared with classical bottom-up approaches. Our online aggregation strategies leverage similarities between shifts in distribution at the local and global scales in order to adapt more quickly. We also propose a new method for designing relevant experts in this context.

1.1. Contributions and outline of the paper

In this study, we propose two methods for leveraging information available at a fine scale to improve prediction at a wider scale based on *feature design* combined using *stacked GAMS and RFs*, and *online aggregation of experts*. These methods are presented in Section 3 and illustrated based on two real-world problems. In Section 4, we demonstrate the application of the first method to the problem of electricity load forecasting at the national level using smart meter data. In Section 5, we illustrate the combination of these methods to obtain adaptive methods for forecasting electricity consumption at the national level during the COVID-19 pandemic period using data available at the regional level. We demonstrate the usefulness of our proposed approach in both cases. Our results indicate that both the stacking of GAMs and RFs, and using features designed based on data at a finer scale lead to improvements in the forecasts at wider scale. Moreover, the use of multi-scale information transfer through aggregation of experts increases the quality of wide-scale forecasts. Strikingly, our results indicate that in the two user cases, the proposed methods can improve wide-scale predictions by using fine-scale predictions, even when no hierarchical constraints are implemented.

To allow the reproducibility of our results, the code and data are provided in the supplementary material: https://drive.google.com/file/d/1hdCEHKpVXt6zoSi7n7xEA0oUKW-_uEeD/view?usp=sharing.

2. Concepts and algorithms

In this section, we describe the different statistical tools used in our transfer learning approach: GAMs, (quantile) RFs, and online aggregation of experts.

2.1. GAMs

GAMs (Wood, 2006) are a simple class of models that model a response as a sum of smooth non-parametric functions of the covariates. Partially linear additive models (PLAM) (Amato, Antoniadis, De Feis, & Goude, 2017) are a special case of generalized additive nonparametric models (GAM) that retain the parsimony and interpretability of linear models, and the flexibility of nonparametric additive regression by allowing a linear component for some predictors, which are assumed to have a strictly linear effect, and an additive structure for other predictors. This choice of both linear and non-parametric components reduces the number of degrees of freedom and mitigates the problem known as the "curse of dimensionality".

Given observations $\{(Y_t, \mathbf{X}_t^{(1)}, \mathbf{X}_t^{(2)})\}_{t=1}^n$, where Y_t is the response at time t, $\mathbf{X}_t^{(1)} = (X_{t,1}^{(1)}, \dots, X_{t,d_1}^{(1)})^T$ and $\mathbf{X}_t^{(2)} = (X_{t,1}^{(2)}, \dots, X_{t,d_2}^{(2)})^T$ are vectors of covariates, the partially linear GAM assumes that

$$Y_t = b + \left(\mathbf{X}_t^{(1)}\right)^T \boldsymbol{\beta} + \sum_{j=1}^{d_2} f_j(X_{t,j}^{(2)}) + \epsilon_t, \quad t = 1, \dots, n, \quad (1)$$

where b is the intercept, $\boldsymbol{\beta}$ is the $d_1 \times 1$ vector of unknown coefficients for linear terms, f_j are unknown nonlinear real valued components, and the ϵ_is are i.i.d. random variables with mean zero and variance σ^2 that are independent of the covariates. In order to ensure that the model is identifiable, we require that the linear covariates are centered and that identifiability conditions $\int f_j(t)dt = 0$, $j = 1, \dots, d_2$ hold. For the sake of simplicity, and with some abuse of definition, we refer to these PLAM models as GAMs and denote as $f_k(X_k)$ the effect of variable X_k, which can be linear or non-parametric.

These models and the procedures for estimation and simultaneous consistent variable selection have been shown to cope with high-level aggregate electricity data in previous studies. In particular, Goude, Nedellec, and Kong (2013) applied these models to consumption by French substations and Fan and Hyndman (2012) demonstrated their suitability for regional load forecasting in Australia. Moreover, they can be applied to efficiently forecast electricity data at different levels of aggregation (Amato, Antoniadis, De Feis, Goude, & Lagache, 2021). In our proposed method, GAMs are trained in R using the mgcv library (Wood, 2017).

2.2. RF and quantile regression forest

RFs are a powerful black box approach for modeling complex regression relationships (see Breiman 2001). The very general model that underlies RF regression assumes that $y_t = h(X_t) + \varepsilon_t$, where g is a generic, non-parametric function, and ε_t is an independent Gaussian noise. Due to the generality of the model, RF requires very little prior knowledge about the problem. RFs are obtained by aggregating an ensemble of base learners generated by applying classification and regression trees (CART, see Breiman, Friedman, Olshen, and Stone 1984) to different subsets of the data obtained by bagging and random sampling of the covariates. An important feature of RFs is that they are easy to use for quantile regression, as shown by Meinshausen and Ridgeway (2006).

In our applications, we use the `ranger()`procedure from the R toolbox `ranger` for the RF fits. The default parameters are used (500 trees, $mtry = \sqrt{p}$, unlimited tree depth). In future studies, these values could be optimized in a more refined manner by combining `ranger` with procedures from the R library `caret`, but at the cost of increasing the CPU time.

2.3. Online aggregation of experts

Online robust aggregation of experts (Cesa-Bianchi & Lugosi, 2006) is a powerful model agnostic approach for time series forecasting that involves combining different forecasts (called experts) according to their past performance in a streaming manner. When expert forecasts of a variable of interest at a finer scale are aggregated to forecast this variable at a wider scale, knowledge transfer occurs between these different scales. Aggregation of experts was recently applied in a forecasting competition (see Farrokhabadi, Browell, Wang, Makonin, Su, and Zareipour 2022), where two of the first three teams (see De Vilmarest and Goude 2022, Ziel 2022) applied this approach to forecast the electricity load consumption during the COVID-19 lockdown in a big city (unknown localization). In this changing context, online aggregation allows adaptation to changes in distribution and tracking the performance of the best expert.

We next provide here a brief description of sequential expert aggregation for forecasting. A complete description of this method was given by Cesa-Bianchi and Lugosi (2006). Sequential expert aggregation assumes that data are observed sequentially, where the target variable (electricity consumption in this study) is assumed to be a bounded sequence $Y_1, \ldots, Y_T \in [0, B], B > 0$ that we want to forecast step by step for every time t. At each time t, N experts provide forecasts of Y_t, denoted by $\left(\hat{Y}_t^1, \ldots, \hat{Y}_t^N\right) \in [0, B]^N$. These experts can be obtained from a statistical model, physical model, or expert advice projection. The aggregation algorithm chooses weights $\hat{p}_{j,t} \in \mathbb{R}^N$, and returns a forecast for Y_t as a weighted average $\hat{Y}_t = \sum_{j=1}^N \hat{p}_{j,t}\hat{Y}_t^j$ of the N forecasts. Y_t is then observed and instance $t+1$ begins. In the following, we consider only convex aggregation (with weights $\hat{p}_{j,t}$ summing to one and in $[0, 1]$).

The performance of experts and aggregation forecasts is evaluated according to a convex loss function. We consider the square loss $\ell_t(x) = (Y_t - x)^2$. At time t, expert k suffers loss $\ell_t(\hat{Y}_t^k) = (Y_t - \hat{Y}_t^k)^2$ and the aggregation $\ell_t(\hat{Y}_t) = (Y_t - \hat{Y}_t)^2$. The aim of expert aggregation is to minimize the total loss $\sum_{t=1}^T (Y_t - \hat{Y}_t)^2$, which can be expressed as:

$$\frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2 \quad \triangleq \quad \frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{Y}_t^*)^2 + R_T,$$

where $\hat{Y}_t^*$ is an oracle that can be viewed as an optimal (unknown before the forecasting run) forecast and R_T is the regret term corresponding to the error suffered by our algorithm relative to the error of the oracle. Some algorithms have been proposed for obtaining low regrets. In the present study, we use the ML-Poly algorithm proposed by Gaillard, Stoltz, and Van Erven (2014) and implemented in the R package `opera` (Gaillard & Goude, 2016). This algorithm tracks the best expert or best convex aggregation of experts by assigning more weight to an expert that generates a low regret. This makes this algorithm particularly interesting because parameter tuning is not required.

3. Hierarchical stacking

In this section, we present our methodological contributions. The first method is based on learning new features using data from the source distribution. These features are then used as inputs in a stacked GAM and RF model. The second method is for designing new aggregation strategies to adaptively forecast variables on a bi-level hierarchy.

3.1. Feature design for stacked GAM and RF

Feature design using the source data. In the following, we assume access to two data sets

$$\mathcal{D}_{\mathcal{T}} = \left(\mathbf{X}_t^{\mathcal{T}}, Y_t^{\mathcal{T}}\right)_{t=1,\ldots,n_{\mathcal{T}}} \quad \text{and} \quad \mathcal{D}_{\mathcal{S}} = \left(\mathbf{X}_t^{\mathcal{S}}, Y_t^{\mathcal{S}}\right)_{t=1,\ldots,n_{\mathcal{S}}},$$

where $\mathcal{D}_{\mathcal{T}}$ is the target data set in the sense that the final objective is to forecast $Y_t^{\mathcal{T}}$, with the underlying distribution $\mathbb{P}^{\mathcal{T}}$. $\mathcal{D}_{\mathcal{S}}$ with the underlying distribution $\mathbb{P}^{\mathcal{S}}$ is an auxiliary source data set that shares some common properties with $\mathcal{D}_{\mathcal{T}}$. We then want to exploit $\mathcal{D}_{\mathcal{S}}$ in order to improve the forecast of $Y_t^{\mathcal{T}}$.

In general, the covariates from the source and target data sets $\mathbf{X}_t^{\mathcal{S}}$ and $\mathbf{X}_t^{\mathcal{T}}$ may belong to spaces with different dimensions, but without loss of generality, we may assume that a subset C of covariates exists that is common to both data sets. In the electricity consumption forecasting setting, these common variables can include calendar variables or meteorological variables (at a finer scale in $\mathcal{D}_{\mathcal{S}}$ and at a wider scale in $\mathcal{D}_{\mathcal{T}}$). Thus, it is natural to assume that these features will have similar effects on the variable of interest Y_t in the target and source data sets. To exploit this idea, we propose learning the effect f_k of a common feature $X_{t,k}$ such that $k \in C$ by using the source data set $\mathcal{D}_{\mathcal{S}}$. In particular, we fit a GAM on the data set $\mathcal{D}_{\mathcal{S}}$ and extract the smooth function f_k corresponding to the effect of covariate $X_{t,k}$. We then use f_k to generate a new feature $f_k(X_{t,k}^{\mathcal{T}})$, which we include in the target data set $\mathcal{D}_{\mathcal{T}}$. When the functions f_k are learned from an auxiliary data set $\mathcal{D}_{\mathcal{S}}$ corresponding to observations at a finer scale, adding the corresponding features to the data set $\mathcal{D}_{\mathcal{T}}$ of observations at the wider scale allows the transfer of knowledge in a hierarchical manner.

It should be noted that we can also use this technique to learn the effects of the covariates directly on the target data set $\mathcal{D}_{\mathcal{T}}$ and apply them as new features in the regression. If we use one type of learner (e.g., GAM) to learn the features and combine these features using a different type of learner (e.g., RF), we can take advantage of both types of learners. This idea motivates the stacking of GAMs and RFs, as presented in the following.

Stacked GAM and RF. Our models for the target problem are obtained by stacking GAMs and the correction provided by RF regression trained on the target data set.

GAMs provide interpretable models and a natural way to incorporate expert knowledge into a statistical model. In addition, due to the smoothness assumptions imposed on GAM functionals, GAMs provide a good representation of the effects of important features and they can extrapolate from training data. However, they only model the influence of pre-specified covariates or pairs of covariates, and thus they may fail to account for some interactions between inputs.

By contrast, RFs can model complex regression relationships (see Breiman 2001), where their black box design can effectively capture complex nonlinear interactions. By definition, RF predictions are restricted to the convex hull of the outcomes Y_t of the training data (all the possible means of y_i and this prevents them from producing aberrant predictions due to extrapolation, even when trained on very small data sets, which is typically the case in a transfer learning framework with a small target data set and a high number of covariates (Balestriero, Pesenti, & LeCun, 2021). To obtain the greatest benefit, we propose stacking these two approaches. Using other black box machine learning methods such as neural nets or boosting trees could obtain potential improvements in future research.

The stacked GAM and RF algorithm (GAM-RF in the following) has the following three steps.

1. First, we fit a GAM model using Eq. (1) on the source data $\mathcal{D}_{\mathcal{S}}$. We use the estimated GAM features $(f_k)_{k\in C}$ to create new features $\left(\left(f_k(\mathbf{X}^{\mathcal{T}}_{k,t})\right)_{k\in C}\right)_{t=1,\ldots,n_{\mathcal{T}}}$ for the target data set.
2. We compute estimates of forecasting residuals on the target data set $\mathcal{D}_{\mathcal{T}}$ (either by cross-validation, block cross-validation, or forecasting errors in an online forecasting setting) denoted by $\widehat{\varepsilon}_t$.
3. We then fit a RF model on the augmented target data set $\left(\widehat{\varepsilon}_t, \mathbf{X}^{\mathcal{T}}_t, \left(f_k(\mathbf{X}^{\mathcal{T}}_{k,t})\right)_{k\in C}\right)_{t=1,\ldots,n_{\mathcal{T}}}$ to predict the GAM residuals $\widehat{\varepsilon}_t$. The final forecasts are obtained by summing the GAM forecasts and the corrections provided by the RF.

The method presented above allows the transfer of information through the new features f_k, which are used as inputs in the RF. In Section 4, we illustrate this methodology by applying the stacked GAM-RF to predict the electricity load at the national level for the United Kingdom by leveraging data available at a finer scale collected by smart meters.

3.2. Online aggregation of stacked experts for a bi-level hierarchy

Our experiments presented in Sections 4 and 5 show that GAM and RF stacking can exploit the data in the source data set to improve prediction on a related target data set. However, this method is based on the assumption that the source and target distributions are constant, and thus it may not be robust if these distributions change. In many hierarchical prediction situations, it is natural to assume that changes in the distribution at the least aggregated level (i.e., on the source data) and at the most aggregated level (i.e., on the target data) are related. We may then want to take advantage of the data available for the source problem to learn these changes more quickly and obtain more adaptive forecasts for the target problem. To achieve this goal, we propose using the online aggregation of quantile experts. We consider a hierarchical forecasting setting where the data y_t are observed at a global level and in K zones $y_{z,t}$ such that $y_t = \sum_{z=1}^{K} y_{z,t}$. We denote y_t^{norm} (resp. $y_{z,t}^{norm}$) as the normalized load at the global (resp. zonal) level. This normalization involves dividing these time series by their empirical mean (computed on the source set). We propose different original methods for creating *experts* that will be aggregated online.

3.2.1. Experts

Online expert aggregation leverages the diversity of predictions made by different experts by combining their predictions. To obtain a diverse set of experts, we train our models to predict different quantiles of the target distribution. Designing experts for low and high quantiles has several advantages. First, if these experts are aggregated online to track the changes in the distribution of the load using convex aggregation, then they are particularly relevant because there is a high probability that the real consumption falls in the convex hull of the quantile experts. Second, this method allows us to obtain experts with similar behavior across regions, which can share weights between the different regions and at the national level. Indeed, it is reasonable to assume that when an expert receives a low weight in one region, it must receive a low weight in all regions. For example, in Section 5, we study the problem of electricity load forecasting in France at the national level using regional data. Measures taken in response to the COVID-19 epidemic in France resulted in a decrease in the electricity load throughout the country and this change will correspond to low-quantile experts for the different regions receiving higher weights in the aggregation. Considering a vectorial aggregation model allows us to take advantage of the similar behavior of quantile experts across regions.

In addition, we increase the diversity of the set of experts by considering different models to predict these quantiles, which we describe as follows. The stacked GAM and RF presented in Section 3.1 can be computed for each zone. We propose two methods for computing the stacking by considering each zone individually (individual GAM-RF) and considering common models for zones and at the global level (common GAM-RF). We describe the experts in the following.

- **GAM:** A GAM is fitted on each zone and at the global level on the normalized data to obtain $K+1$ scaled experts.
- **Individual GAM-RF:** Using quantile regression forests, for each zone and at the global level, we fit five RFs on the residuals of the GAM model. These RFs predict quantiles at levels 0.05, 0.1, 0.5, 0.9, and 0.95. By stacking the predictions of these RFs and the GAM, we obtain five scaled experts for each zone.
- **Common GAM-RF:** Using quantile regression forests, we fit five RFs on the aggregated residuals for all zones and the global level. These RFs predict quantiles at levels 0.05, 0.1, 0.5, 0.9, and 0.95. By stacking the predictions of these RFs and the GAM, we obtain five scaled experts for each zone.

These approaches are illustrated in Fig. 1. Quantile experts can be considered as possible scenarios for the evolution of the data distribution that we try to track online in the aggregation. Common RFs are used to improve the transfer efficiency and capture common dynamics between the zones.

3.2.2. Aggregation strategies

The strategy described in Section 3.2.1 yields 11 experts for each one of K zones and for the global level: one GAM expert, five GAM-RF stacked experts trained zone by zone, and five GAM-RF stacked experts trained on the aggregated data. Thus, we obtain $11 \times K$ experts. To combine the predictions of these experts, we propose four aggregation strategies that consider the hierarchical structure of the data in different ways. The algorithms are described below and illustrated in Figs. 2 and 3.

- **Fully disaggregated model**: We use the full set of $11(K+1)$ scaled forecasts as experts and the scaled global response y^{norm} as our target variable. The prediction is then multiplied by the average value of the response at the global level.

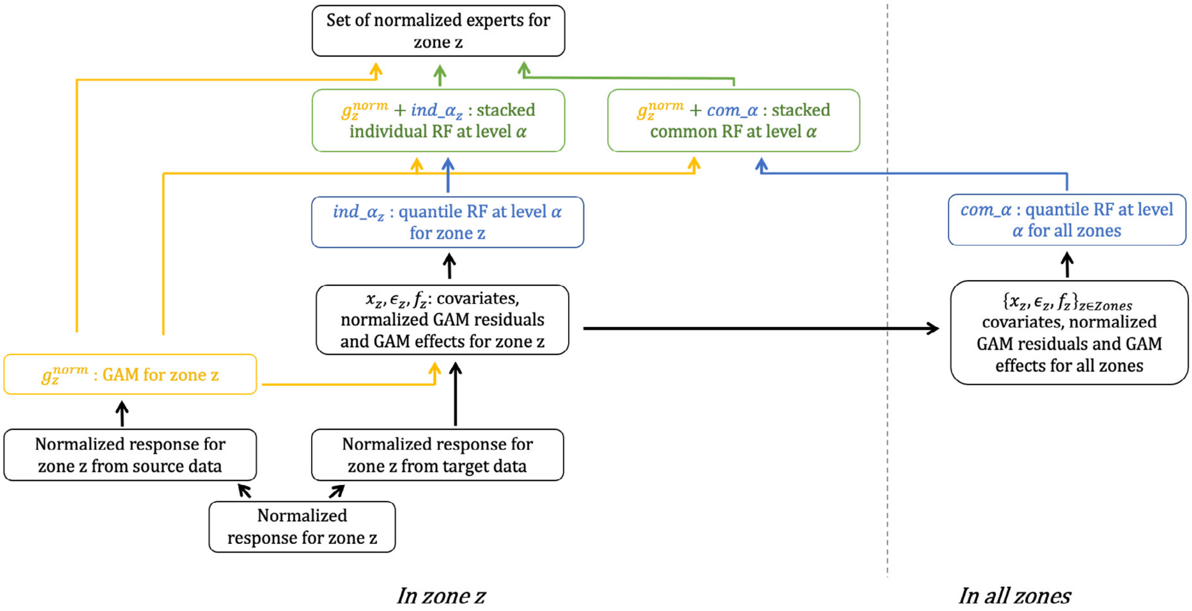

Set of normalized experts for zone z
g_z^{norm} + ind_α_z : stacked individual RF at level α
g_z^{norm} + com_α : stacked common RF at level α
ind_α_z : quantile RF at level α for zone z
com_α : quantile RF at level α for all zones
x_z, ϵ_z, f_z: covariates, normalized GAM residuals and GAM effects for zone z
$\{x_z, \epsilon_z, f_z\}_{z \in Zones}$ covariates, normalized GAM residuals and GAM effects for all zones
g_z^{norm} : GAM for zone z
Normalized response for zone z from source data
Normalized response for zone z from target data
Normalized response for zone z
In zone z
In all zones

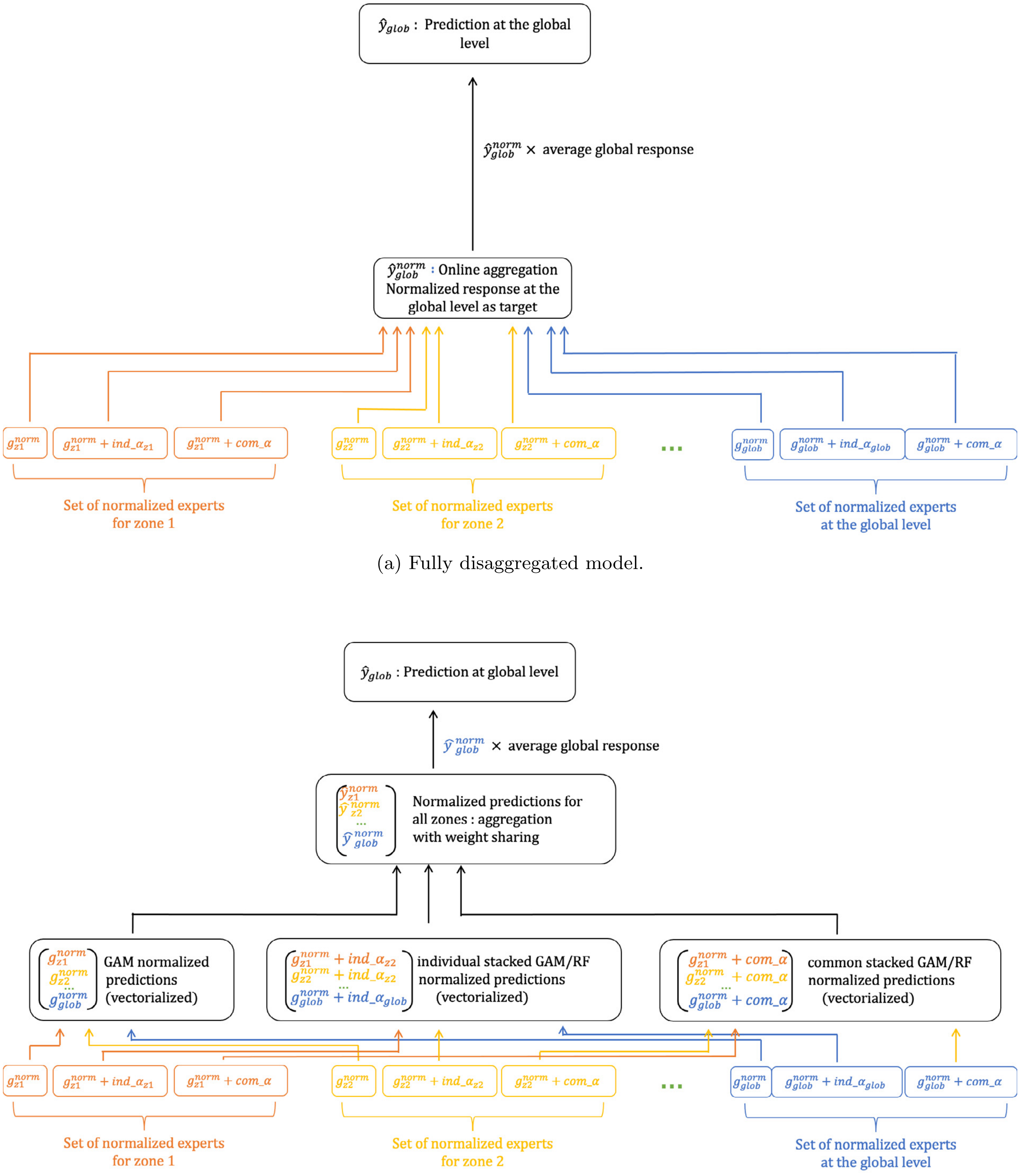
Fig. 1. Experts used for predicting normalized responses for the different zones and at the global level.

- **Vectorial aggregation**: We illustrate the possibility of sharing weights between the zones and at the global level. We aim to predict the time series of the (K+1)-dimensional vector corresponding to the scaled response in each zone and at the global level. In particular, we aggregate 11 vectorial, (K+1)-dimensional experts corresponding to the predictions of the GAMs and of the 10 stacked GAM-RF. The prediction corresponding to the global level is then multiplied by the average value of the response at the global level to forecast y.
- **Hierarchical aggregation, scaled predictions**: First, we aggregate the 11 experts in each zone using the scaled response for a zone z, y_z^{norm} as a target and obtain K experts. Next, we aggregate these K experts and the quantile experts at the global level, with the scaled global response y^{norm} as our target variable. The prediction is then multiplied by the average value of the response at the global level.
- **Hierarchical aggregation, unscaled predictions**: We again aggregate the 11 experts in each zone using the scaled response for the corresponding zone as a target and obtain K experts predicting the normalized response at the zonal level. Next, we multiply their predictions by the average value of the response for the corresponding zone y_z and sum these predictions in order to obtain a forecast of the global level y. We note that this aggregation method is the only one that enforces the coherency between the prediction at the zonal level and the prediction at the global level.

In Section 5, we apply this methodology to adaptively predict the electricity load at the national level in France during the first COVID-19 lockdown using data at the regional level. We note that our objective is the forecast at the national scale and the forecasts at the regional scale are only used to improve this aggregated forecast. Thus, we do not require that the forecasts are coherent (the sum of the forecasts at the regional level does not necessarily equal the forecast at the national level). In fact, the experiments presented in Section 5 indicate that aggregation methods that do not respect coherency (such as vectorial aggregation) can obtain more accurate results than methods that respect coherency (i.e., hierarchical aggregation with unscaled predictions).

4. Transfer learning for forecasting aggregated smart meter data

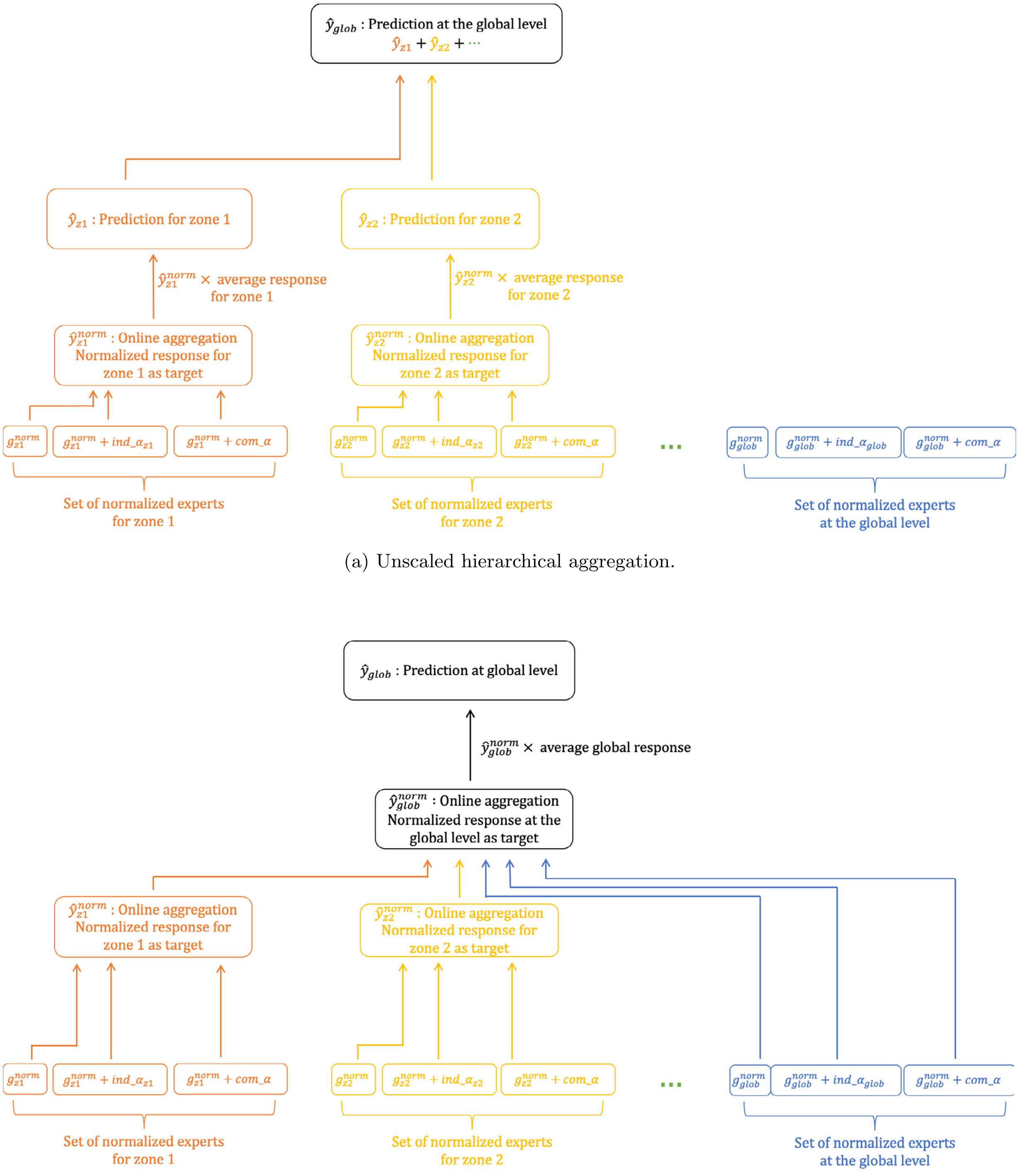
In this section, we illustrate the methodology using a data set that is commonly used for the calibration of electricity consumption forecasting models. The data set comprises aggregate semi-hourly consumption data for the national load in the United Kingdom, and observations of some meteorological and calendar variables. Our goal is to forecast the electricity consumption at the national level from December 2009 to August 2010 (this period is referred to as the test set). We assume that access is available to data at the national level covering the period from April 2005 to November 2009 (the learning set) and data from smart meters for a shorter period (from April 2009 to August 2010). In this first example, we compare the performance of GAM, RF, and stacked GAM-RF trained using data at the national scale, and the performance of stacked GAM-RF using features learned from smart meter data. This comparison allows us to highlight the advantage of stacking GAM and RF, and of transferring the GAM features learned at the finer scale, and to decompose the contributions due to stacking GAM and RF, and to using these new features.


$\hat{y}_{glob}$: Prediction at the global level
$\hat{y}_{glob}^{norm}$ × average global response
$\hat{y}_{glob}^{norm}$: Online aggregation Normalized response at the global level as target
g_{z1}^{norm}
$g_{z1}^{norm} + ind\_\alpha_{z1}$
$g_{z1}^{norm} + com\_\alpha$
g_{z2}^{norm}
$g_{z2}^{norm} + ind\_\alpha_{z2}$
$g_{z2}^{norm} + com\_\alpha$
...
g_{glob}^{norm}
$g_{glob}^{norm} + ind\_\alpha_{glob}$
$g_{glob}^{norm} + com\_\alpha$
Set of normalized experts for zone 1
Set of normalized experts for zone 2
Set of normalized experts at the global level
$\hat{y}_{glob}$: Prediction at global level
$\hat{y}_{glob}^{norm}$ × average global response
$\hat{y}_{z1}^{norm}$, $\hat{y}_{z2}^{norm}$, ..., $\hat{y}_{glob}^{norm}$
Normalized predictions for all zones : aggregation with weight sharing
g_{z1}^{norm}, g_{z2}^{norm}, ..., g_{glob}^{norm}
GAM normalized predictions (vectorialized)
$g_{z1}^{norm} + ind\_\alpha_{z2}$, $g_{z2}^{norm} + ind\_\alpha_{z2}$, ..., $g_{glob}^{norm} + ind\_\alpha_{glob}$
individual stacked GAM/RF normalized predictions (vectorialized)
$g_{z1}^{norm} + com\_\alpha$, $g_{z2}^{norm} + com\_\alpha$, ..., $g_{glob}^{norm} + com\_\alpha$
common stacked GAM/RF normalized predictions (vectorialized)
g_{z1}^{norm}
$g_{z1}^{norm} + ind\_\alpha_{z1}$
$g_{z1}^{norm} + com\_\alpha$
g_{z2}^{norm}
$g_{z2}^{norm} + ind\_\alpha_{z2}$
$g_{z2}^{norm} + com\_\alpha$
...
g_{glob}^{norm}
$g_{glob}^{norm} + ind\_\alpha_{glob}$
$g_{glob}^{norm} + com\_\alpha$
Set of normalized experts for zone 1
Set of normalized experts for zone 2
Set of normalized experts at the global level

(a) Fully disaggregated model.

(b) Vectorial aggregation.

Fig. 2. Fully disaggregated and vectorial aggregation strategies.


$\hat{y}_{glob}$: Prediction at the global level
$\hat{y}_{z1} + \hat{y}_{z2} + \cdots$
$\hat{y}_{z1}$: Prediction for zone 1
$\hat{y}_{z2}$: Prediction for zone 2
$\hat{y}^{norm}_{z1}$ × average response for zone 1
$\hat{y}^{norm}_{z2}$ × average response for zone 2
$\hat{y}^{norm}_{z1}$: Online aggregation Normalized response for zone 1 as target
$\hat{y}^{norm}_{z2}$: Online aggregation Normalized response for zone 2 as target
g^{norm}_{z1}
$g^{norm}_{z1} + ind\_\alpha_{z1}$
$g^{norm}_{z1} + com\_\alpha$
g^{norm}_{z2}
$g^{norm}_{z2} + ind\_\alpha_{z2}$
$g^{norm}_{z2} + com\_\alpha$
...
g^{norm}_{glob}
$g^{norm}_{glob} + ind\_\alpha_{glob}$
$g^{norm}_{glob} + com\_\alpha$
Set of normalized experts for zone 1
Set of normalized experts for zone 2
Set of normalized experts at the global level
$\hat{y}_{glob}$: Prediction at global level
$\hat{y}^{norm}_{glob}$ × average global response
$\hat{y}^{norm}_{glob}$: Online aggregation Normalized response at the global level as target
$\hat{y}^{norm}_{z1}$: Online aggregation Normalized response for zone 1 as target
$\hat{y}^{norm}_{z2}$: Online aggregation Normalized response for zone 2 as target
g^{norm}_{z1}
$g^{norm}_{z1} + ind\_\alpha_{z1}$
$g^{norm}_{z1} + com\_\alpha$
g^{norm}_{z2}
$g^{norm}_{z2} + ind\_\alpha_{z2}$
$g^{norm}_{z2} + com\_\alpha$
...
g^{norm}_{glob}
$g^{norm}_{glob} + ind\_\alpha_{glob}$
$g^{norm}_{glob} + com\_\alpha$
Set of normalized experts for zone 1
Set of normalized experts for zone 2
Set of normalized experts at the global level

(a) Unscaled hierarchical aggregation.

(b) Scaled hierarchical aggregation.

Fig. 3. Unscaled and scaled hierarchical aggregation strategies.

Table 1
Variables used in models (2) and (3).

Variable name	Description
Y_t^c	de-trended electricity load
Temp_t	weighted temperature
Temp99_t	weighted and exponentially smoothed temperature
Instant_t	instant in the day
DayType_t	day of the week
Holiday_t	binary variable indicating public holidays
LongWeekEnd_t	binary variable indicating the presence of a long weekend
ToY_t	time of year

4.1. Data

4.1.1. National data

This UK national semi-hourly electricity consumption data set is provided by the European Grid Standards Office (see https://www.nationalgrideso.com/balancing-data/data-finder-and-explorer) and it covers the period between April 2005 and December 2010. We add features comprising temperature data obtained from the National Oceanic and Atmospheric Administration (NOAA)[1] for the ten largest cities in the UK: London, Birmingham, Glasgow, Sheffield, Bradford, Liverpool, Edinburgh, Manchester, and Bristol. We then compute a weighted average T_t of the temperatures recorded in these ten stations with weights proportional to the official population of each city, and finally perform exponential smoothing of this weighted average with the parameters 0.2, 0.05, and 0.01.

4.1.2. Data from smart meters

This data set corresponds to data obtained from smart meters at an individual scale in the UK. This data set was obtained in the Energy Demand Research Project launched by Ofgem on behalf of the UK Government in 2007 (see AECOM 2018, Schellong 2011[2]), where power consumption for approximately 60,000 households was collected at half hourly intervals for about two years. We consider a subset of 1925 customers from April 2009 to August 2010 located in two regions of the UK: southeast (around Brighton) and northwest (around Glasgow). We consider temperatures in each region obtained from the NOAA. To this data set, we add supplementary calendar covariates, such as the time of year, day type, and sunrise and sunset times throughout the year.

4.2. Models and forecasting

The fitting procedure used to forecast electricity consumption at the national level can be described as follows.

We note a trend in the consumption time series over the period from April 2005 to August 2010. We estimate this trend in a very simple way by fitting a nonparametric Gaussian model $Y_t = \mu + s(t) + \varepsilon_t$ to the series of observations, where the trend $s(t)$ is represented in a cubic spline function base and the number of knots is limited to three. In the following, we subtract this trend and aim to forecast the national *de-trended* consumption, which is then given by $Y_t^c = Y_t - \widehat{s}(t) - \hat{\mu}$.

We apply the stacked GAM and RF methodology to predict the national load consumption using only data available at the national level. Note that this is a special case of the general transfer learning framework with $\mathcal{D}_\mathcal{T} = \mathcal{D}_\mathcal{S}$, where the final forecast is obtained using RF based on the data enriched with the transfer of information performed using the design of new features. Similar GAMs have already been used successfully to forecast electrical loads at the smart meter resolution (Gilbert, Browell, & Stephen, 2023). Fasiolo, Wood, Zaffran, Nedellec, and Goude (2021) used similar GAMs (but their quantile version was used for probabilistic forecasting) at the national level in the UK. First, we fit a semi-parametric GAM on the learning set. The GAM is given by

$$\begin{aligned} Y_t^c &= \textstyle\sum_{j=1}^7 m_j \mathbf{I}_{\text{DayType}_t=j} + m_8 \mathbf{I}_{\text{Holiday}_t=1} \\ &+ m_9 \mathbf{I}_{\text{LongWeekEnd}_t=1} \\ &+ g_1(\text{Instant}_t, \text{Temp}_t) + \textstyle\sum_{j=1}^7 f_j(\text{Instant}_t) \mathbf{I}_{\text{Daytype}_t=j} \\ &+ s(\text{ToY}_t) \\ &+ s(\text{Temp99}_t) + \varepsilon_t \end{aligned} \tag{2}$$

where the variables are presented in Table 1 and ε_t is a centered Gaussian noise. Each univariate smooth component of the GAM model above is fitted using regression spline functions with 40 knots (50 knots for ToY) and a tensor basis of spline functions for the interaction between time and temperature with 20 and 10 knots, respectively.

Next, after fitting (2), we extract the estimated effects g_1 and f_j of the features, and add them to the set of initial covariates. This enriched data set is then used to train the RF. We note that the initial number of covariates in the database considered in Eq. (2) is seven and the number of additive components extracted by using the GAM methodology is 15. Thus, after transfer, the number of observed covariates in the sample is equal to 22. We apply the GAM-RF stacking methodology to fit a nonparametric regression called `GAM.RF.nat` on the training sample. Finally, we evaluate the prediction of this stacked GAM and RF on the test sample, and compare it to the predictions obtained with the GAM (2) model. We also compare this model to a standard regression model by RFs denoted as `RF.nat`.

[1] https://www.noaa.gov/

[2] https://www.ofgem.gov.uk/gas/retail-market/metering/transition-smart-meters/energy-demand-research-project

Table 2
Errors in predictions for the learners GAM.nat, RF.nat, GAM.RF.nat, and GAM.RF.local.

	GAM.nat	RF.nat	GAM.RF.nat	GAM.RF.local
Root mean squared error	1409 MW	1339 MW	1214 MW	1193 MW
Mean absolute percentage error	2.670	2.560	2.360	2.310
Number of covariates	7	7	22	32

For a second time, we apply the stacked GAM and RF methods to transfer information from the smart meters data. We begin by computing the total consumption by customers based on the smart meter data set and fit a GAM model to forecast this total. Using the same methodology applied to the national data, we obtain the model presented in Eq. (3). We use a simpler model than the national model because the data set used to train it is smaller.

$$\begin{aligned} y_t \quad = \quad & \textstyle\sum_{j=1}^{7} m_j \mathbf{I}_{DayType_t=j} \\ + \quad & g_1(\mathrm{Instant}_t, T_t) \; + \textstyle\sum_{j=1}^{7} f_j(\mathrm{Instant}_t)\mathbf{I}_{DayType_t=j} \\ + \quad & s(\mathrm{ToY}_t) \\ + \quad & \varepsilon_t \end{aligned} \tag{3}$$

We then extract the ten nonlinear features of this model as supplementary covariates and add them to the data set comprising all original covariates and the effects extracted from the GAM at the national level. Finally, we use these covariates to train the stacked GAM-RF and obtain a model called GAM.RF.local. We note that the GAM (3) used to model the aggregated smart meter load is fitted on the small data set consisting of smart meter data. However, the effects f_k extracted from this model are then used to create new features for each entry of the large national data set. Thus, the RF.nat, GAM.nat, GAM.RF.nat, and GAM.RF.local models are trained on the same number of observations from the national data set.

Determination of the importance by permutation analysis for the variables used in the stacked RF after learning by transfer retains the instant of the day as the most important for GAM.RF.nat, followed by three terms from the national GAM modeling. For the model GAM.RF.local, the most important variables are the instant of the day, followed by two terms from the national GAM and one term from the local GAM.

By analyzing the mean absolute percentage error (MAPE) and the root mean squared error (RMSE) values for the different methods, as presented in Table 2, we see that for the UK national data set, the RFs are more efficient than the adopted reference model GAM. Interestingly, the stacked GAM and RF trained using only national data GAM.RF.nat outperforms these two models. This finding indicates that the stacking of GAM and RF allows us to obtain the greatest benefit, where the RF can correct for the effects or interactions between variables (such as the instant of the day) that are not captured well by the GAM, as well as being robust to the large number of covariates used as inputs (up to 28). Finally, the best model in both terms of the MAPE and RMSE values is obtained by stacking GAM and RF using the effects learned from both national and smart meter data. These results highlight the value of leveraging data available at a finer scale even when no hierarchical constraints are implemented in the algorithm.

5. Electricity load forecasting during the first COVID-19 lockdown

In this section, we apply our methodology to short-term electricity load forecasting during the COVID-19 lockdown and post-lockdown period in France at a resolution of half an hour and at the national level. In particular, we leverage information available at the regional level. Electricity consumption was significantly affected by the measures taken by the government to cope with the epidemic because closures of non-essential businesses and stay-at-home directives decreased the electricity consumption by about 10%, and changes occurred in the daily and weekly patterns (see Obst, de Vilmarest and Goude 2021, for a description of the impacts of these measures on electricity consumption in France). Common models trained on historical data, which rely on calendar and weather data, fail to account for these significant changes. Similarly, transfer learning methods that rely on data present at a finer (e.g., regional) scale will make poor predictions if trained on data with a different distribution to that of the target, especially if the relationships between local and global variables change over time. Thus, these models trained on data from the pre-pandemic period make relatively large prediction errors on the period following the start of the lockdown. To ensure the adaptativity of our models, we combine the stacked GAM-RF methodology presented in Section 3.1 with the online aggregation of quantile experts presented in Section 3.2.2.

Transfer learning has been shown to be essential for addressing the problem of electricity load forecasting during the COVID-19 pandemic. Data for this period are scarce, especially since we want to make predictions from the very beginning, so it is crucial to use information from the pre-pandemic period to predict power consumption during the pandemic period. In particular, we use the methods presented above to transfer information from the large data set corresponding to historical electricity consumption during the pre-pandemic period (source period) to improve predictions during the pandemic period (target period), which is again conducted using the stacked GAM and RF. This transfer learning algorithm allows us to rely on a GAM trained on a large set of observations of historical electricity consumption from the source distribution, as well as correcting for the error on the target using RF based on scarce observations.

In addition, due to the important changes in electricity consumption following the lockdown, we expect that the relationships between effects learned on regional data and national load will also change. Indeed, our studies indicate that containment measures induce changes in electricity consumption at the regional level, but they differ according to the region considered. To utilize electricity consumption data available at the regional level, our method must remain adaptive to changes in the distributions of both national and regional data, which is achieved by the online aggregation of experts, thereby allowing us to combine forecasts at the regional and national levels in an adaptive manner. We forecast the electricity consumption separately region by region using stacked GAM and RF, and then combine the forecasts of these regional models in order to predict the national electricity consumption in a hierarchical manner. The hierarchical model captures regional phenomena that are not apparent at a more aggregated scale and leverages this information to improve predictions at the national level. Thus, our methods allow the transfer of knowledge at both a temporal level (data from the pre-pandemic period are used to improve forecasts during the pandemic period) and hierarchical level (regional predictions are used to produce forecasts at the national level).

The remainder of this section is organized as follows. In Section 5.1, we present the data used to design and evaluate our models. In Section 5.2, we present the models used for forecasting the electricity consumption at national and regional levels. The results obtained in our study are presented in Section 5.3. First, we compare the performance of different approaches, before presenting a more detailed analysis of the stacked GAM and RF, and the online aggregation of experts.

5.1. Data

The data from the French Transmission System Operator (TSO, RTE) comprise electricity consumption (in MW) at a half-hourly temporal resolution at the French national level (“Load") and for the 12 metropolitan administrative regions (not including Corsica): Nouvelle Aquitaine, Auvergne Rhônes-Alpes, Bourgogne-Franche-Comté, Occitanie, Hauts-de-France, Normandie, Bretagne, Centre-Val de Loire, Île-de-France, Pays de la Loire, Provence-Alpes-Côte d’Azur, and Grand Est. Our goal is to forecast the French national consumption using the regional loads information. For all load consumption data, we compute the lags for one day and one week, and denote then by the subscripts“0.48" and “0.336”, respectively.

Our models use the temperature and weighted temperature as explanatory variables. These variables were collected from the website of the French weather forecaster M\'et\'eoFrance. For each region, we compute the weighted mean of meteorological stations where the weights are proportional to $\exp(-dist)$, and $dist$ is the distance of the station to the barycenter of each region. Note that we use the observed temperatures instead of their predicted values in our forecast. Thus, we cancel out the errors caused by the uncertainty of a particular weather forecast, which allows more precise comparison of the different models. Moreover, this choice allows us to only use open data to ensure the reproducibility of our results.

Our models also rely on variables that characterize the impacts of the restrictions implemented to fight the epidemic. The first of these variables is the Oxford COVID-19 Government Response Tracker. This index (freely available at https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker) aggregates indicators that characterize the measures taken by governments to mitigate the epidemic in terms of containment, health, and economic support. This index is available at the national level. The methodology used to calculate the index and the measures on which it is based are known a few days in advance, so we assume that it is known for the day that we wish to forecast. The remaining variables used to characterize the impacts of lockdown measures are Google Mobility Indices. These indices are provided by Google and obtained by aggregating geolocalization data. They characterize changes in the frequentation of categorized places (residential, workplaces, transport, parks, grocery and pharmacy, retail, and recreation). The data are freely available (at https://www.google.com/covid19/mobility/) but with a delay of slightly less than a week. Therefore, we consider lagged versions of these indicators in our prediction. The government response and mobility indices are available from January and February 2020, respectively. Therefore, we do not use them as covariates in the source model, but only in the target model.

In the following, our source models are the models trained on historical data collected between the beginning of 2012 and the end of August 2019. We evaluate their performance on data with the same distribution during the pre-lockdown period ranging from September 2019 to March 15, 2020, and compare their performance on data from the target distribution ranging from March 16 to September 17. By contrast, the models specific to the lockdown and post-lockdown period, which are the target models in the following, are retrained every day during this target period to exploit all of the available observations. It should be noted that the first lockdown in France officially started on March 17, but we consider March 16 as the first day of the target distribution because the electricity consumption pattern had already changed by that day.

5.2. Models

5.2.1. GAMs for the pre-pandemic period

We use GAMs to predict the electricity load under normal circumstances. We fit one model for each region of mainland France, as well as one at the national level to obtain 13 models. To consider the daily electricity consumption patterns, each model comprises 48 GAMs fitted independently and the electricity load is forecast at a given instant on the day. Thus, the 624 time-series corresponding to the 48 half-hours in the 12 regions and the national level are treated independently. In order to compare the predictions, terms, and errors of the models, the regional and national electricity loads are normalized by dividing by the average value for the region and for the half-hour considered. GAMs are then fitted to predict this normalized load. In the following, we denote y and y^{norm} as the load and *normalized* load, respectively.

Table 3
Variables at time t used in model (4).

Variable name	Description
$y_{z,t}^{norm}$	normalized electricity load for zone z
$\mathrm{Daytype}_t$	categorical variable indicating the day of the week
DLS_t	binary variable indicating whether t is in summer hours or winter hours
ToY_t	time of year
$\mathrm{Temp}_{z,t}$	temperature in zone z
$\mathrm{Temp95}_{z,t}$	weighted and exponentially smoothed temperature of smoothing factor 0.95
$\mathrm{Temp99}_{z,t}$	weighted and exponentially smoothed temperature of smoothing factor 0.99
$\mathrm{TempMin99}_{z,t}$	minimal value over the day of $\mathrm{Temp99}_{z,t}$
$\mathrm{TempMax99}_{z,t}$	maximal value over the day of $\mathrm{Temp99}_{z,t}$
$\mathrm{Load.48}_{z,t}$	normalized load on the day before in zone z
$\mathrm{Load.336}_{z,t}$	normalized load in the week before in zone z

The model used to predict the electrical load for zone z at time t corresponding to the hth half-hour of the day is as follows:

$$\begin{aligned} y_{z,t}^{norm} = & \sum_{i=1}^{7}\sum_{j=0}^{1}\alpha_{i,j}^{(z,h)}\mathbb{1}_{\mathrm{DayType}_t=i}\mathbb{1}_{\mathrm{DLS}_t=j} \\ & + \sum_{i=1}^{7}\beta_i^{(z,h)}\mathrm{Load.48}_{z,t}\mathbb{1}_{DayType=i} + \gamma^{(z,h)}\mathrm{Load.336}_{z,t} \\ & + f_1^{(z,h)}(t) + f_2^{(z,h)}(\mathrm{ToY}_t) + f_3^{(z,h)}(t, \mathrm{Temp}_{z,t}) \\ & + f_4^{(z,h)}(\mathrm{Temp95}_{z,t}) + f_5^{(z,h)}(\mathrm{Temp99}_{z,t}) \\ & + f_6^{(z,h)}(\mathrm{TempMin99}_{z,t}, \mathrm{TempMin99}_{z,t}) + \varepsilon_{z,t} \qquad (4) \end{aligned}$$

where $\varepsilon_{z,t}$ is Gaussian white noise and the variables are presented in Table 3. Each univariate smooth component of the GAM model above is fitted using regression spline functions with 20 knots for ToY, ten knots for Temp95 and Temp99, 5 knots for Date, and a tensor basis of spline functions for the interaction between time and temperature with three and five knots, respectively.

5.2.2. Quantile GAM-RF expert aggregation

We design experts by stacking GAM and RF according to the methodology described in Section 3. The RFs are trained in a streaming manner on the target data (pandemic and post-pandemic period). Based on the results given in Section 4, we apply the usual covariates as the inputs for these RFs but also the GAM effects learned on the source data set. Interestingly, preliminary results given in Appendix A indicate that although the RF inputs are high-dimensional, variable selection only marginally affects the performance of our model. The RFs appear to be robust against the high dimensionality of the features, even in the early days of lockdown when few observations are available.

Using RFs to correct the errors of the GAM during the pandemic period allows us to obtain an adaptive model that can produce predictions from the very beginning of the target period. It should be noted that the corrections of the RF remain small compared with the predictions of the GAM, where the first order of the prediction is given by the source model trained on the large set of historical data, and the corrections learned on the scarce observations from the target data set only provide a second order correction.

For each of the 12 regions and at the national level, we obtain 11 experts corresponding to the GAM experts, the five GAM-RF quantiles experts trained on the residuals of the zone, and the five GAM-RF quantiles experts trained on the aggregated residuals. We then compare four aggregation techniques (full disaggregated model, vectorial aggregation, hierarchical aggregation of scaled predictions, and hierarchical aggregation of unscaled predictions).

5.3. Results

In the following, we compare the methods presented above. First, we compare their performance in terms of the MAPE and RMSE values in Section 5.3.1. We then present the importance according to permutation analysis of the RFs in Section 5.2.2. Finally, we analyze and compare the different aggregation methods in Appendix B.2.

5.3.1. Performance

Table 4 compare the MAPE and the RMSE values for the four methods by using GAM at the national level and the stacked individuals RF for predicting the median of the residuals at the national level.

We split the test period into three sub-periods. In the pre-pandemic period between September 1, 2019 and March 15, 2020, only the GAM predictions are available. During this period, vectorial aggregation is not applicable because there is only one type of expert. The lockdown period ranges between March 16 and May 11, and training data are very scarce during this period, and models must adapt rapidly to dramatic changes in the electricity consumption patterns. The post-lockdown period from May 12 to September 17 corresponds to a new change in the load pattern due to a relative rebound in activity, to which the models must adapt.

GAM-RF stacking improves GAM significantly. Using stacked RFs to predict the median of the GAM residuals at the national level is sufficient to decrease MAPE during and after the lockdown by 50% and 45%, respectively. Except for vector aggregation, all hierarchical aggregation strategies outperform GAM and GAM-RF during the lockdown period, and these results indicate that online aggregation is an efficient method for considering information available at a finer scale. Our analysis in Appendix B shows that the average error with regional GAM is much larger than that with GAM at the national level due to the larger fluctuations present at the finer scale. Interestingly, aggregating these low-accuracy models obtains

Table 4
Mean absolute percentage error and root mean squared error for the stacked GAM and RF models.

Model	2019/09/01-2020/03/15	2020/03/16-2020/05/11	2020/05/12-2020/09/17
GAM	1.36%, 1030 MW	4.82%, 2838 MW	1.84%, 1045 MW
Individual stacked GAM-RF	Non applicable	2.41%, 1813 MW	1.03%, 592 MW
Full disaggregated	1.20%, 910 MW	2.26%, 1716 MW	1.09%, 609 MW
Hierarchical aggregation scaled	1.14%, 861 MW	2.21%, 1648 MW	1.07%, 609 MW
Hierarchical aggregation unscaled	1.20%, 907	2.08%, 1553 MW	1.02%, 593 MW
Vectorial aggregation	Non applicable	2.56%, 1885 MW	0.91%, 521 MW

better performance than GAM at the national level, even in the pre-pandemic period. This confirms the usefulness of aggregating quantile GAM-RF experts to track changes in the data. Vectorial aggregation performs rather poorly compared with other aggregation strategies during the lockdown but better than all other models after the end of the lockdown.

In Appendix B, we analyze the stacked GAM and RF. We plot the evolution of the importance of the variables over time. Our results show that variables important for predicting one quantile tend to be important for predicting the other quantiles. Moreover, the effects of GAM are among the most important covariates for predicting the GAM residuals. Using these effects as covariates allows the transfer of information about the impacts of weather and calendar variables learned on the large data set of pre-pandemic observations. We also note that as time passes and the size of the training set for the RF increases, relevant variables such as the Government Response Tracker or relative occupation of some places of interest become more important for prediction. Conversely, spurious variables are discarded as unimportant. Interestingly, the common RFs trained on residuals across all regions detect these relevant variables more quickly than the individual RFs trained solely on residuals at the national level. This highlights the benefit of aggregating data across zones and scales in this sparse data context.

The fact that scaled and unscaled hierarchical aggregation perform similarly is somewhat counterintuitive given that in the scaled model, aggregation must learn the contributions of the different regions to national consumption. To analyze this phenomenon, Appendix B shows the relative weights given to the experts corresponding to the different regions. We find that the weights in the unscaled hierarchical aggregation do not correspond to the proportions of electricity consumed by the regions, and that they typically exhibit much more flexibility than the true weights. The fact that scaled hierarchical aggregation outperforms its unscaled counterpart both in the pre-pandemic and in the post-lockdown period suggests that the flexibility provided by the second layer of aggregation used in the scaled model compensates for the lack of knowledge of the relative contributions of the different regions.

We also consider the weights given by the experts in vectorial aggregation and note that it gives the highest weight to the GAM and median stacked RF experts, which appear to be the most relevant experts across all regions; however, these weights are highly unstable during the beginning of the lockdown. The performance of vectorial aggregation during this period is worse than that of all other aggregation models, and also that of the stacked RFs for predicting the median of residuals. This behavior reflects the fact that the impact of the pandemic differs greatly among regions, as shown in Appendix B. By contrast, vectorial aggregation performs best during the post-lockdown period and it appears to be a promising approach for predicting consumption under normal circumstances.

6. Conclusions and future work

We propose new transfer learning methods designed for forecasting time series observed at different hierarchical scales. We present two different settings and illustrate them with two different user cases.

1. To transfer information from finer scale (an aggregate of smart meters) to wider scale (national) data when the distribution of the data is stable with time, we propose stacking features from GAMs obtained at these two scales into RFs.
2. To transfer information from local to global data when the distribution of the data changes over time, we propose the hierarchical online aggregation of experts, where the experts are generated at a finer scale (regional level) using quantile stacked RFs.

We demonstrate the utility of our proposed approach in two user cases. In both cases, transfer learning by RF stacking at a single scale significantly improves the forecasting performance of a single GAM or RF model, with an improvement of 14% over GAM and 9% over RF for case one, and 38% over GAM for case two. This supports our original intuition that stacked RFs exploit the ability of GAM to extrapolate and RFs to automatically model interactions between covariates.

Our results are also convincing regarding multi-scale transfer performance. In case one, the day-ahead forecasting performance of the wider scale stacked GAM-RF improves by about 1.5% with our multi-scale transfer algorithm. In case two, the best hierarchical aggregation algorithm is improved by about 10% with the stacked GAM-RF at a wider scale. Our relatively simple strategy for re-scaling plus aggregation performs well in this bi-level hierarchy. In addition, introducing strong constraints on the aggregation weight (vectorial aggregation) may

Table A.5
Mean absolute percentage error for the stacked GAM and RF models.

Selection method	2020/03/16-2020/05/11	2020/05/11-2020/09/17
Lasso with aggregation	2.41%	1.06%
Boruta	2.41%	1.03%
Full model	2.41%	1.03%

be a useful transfer strategy when the experts behave similarly at different scales of the hierarchy, which is the case during the post-COVID-19 period but not during the hard lockdown in France from March–April 2020 (we suspect that the effect of COVID-19 on the electricity load could impact the different regions in an unsynchronized manner).

The main learner used in our method for final forecasting is based on stacked GAM-RF. We could select other machine learning methods, such as tree-based gradient boosting or neural networks, which can be tested in future research. According to our experiments, automatic variable selection when forecasting does not obtain any improvements. However, in a high dimensional setting with a large number of features generated when learning the source, we consider that a possible approach for exploration involves using a regression-reinforced RF (RFRF) approach for forecasting, which may obtain better predictions than RFs. The idea behind RFRF involves exploiting the strength of penalized parametric regression to improve RFs. For example, for RFRFs, we may run smoothly clipped absolute deviations (SCAD) (or LASSO) (see Fan and Li 2001) based selection before RF, and then construct a RF on the residuals from the SCAD (or LASSO) penalized fit. Preliminary simulation results show that RFRFs can exploit the strength of both parametric and nonparametric methods, and they may give reliable predictions in high-dimensional extrapolation problems such as those encountered in transfer learning.

In our first user case, the clustering of smart meter data to generate diverse GAM features is not investigated but this is clearly a possible improvement. Introducing hierarchical constraints on the weights, as proposed by Brégère and Huard (2022), is another potential option for our second user case. Identifying the good warping of weight constraints for vectorial aggregation could also improve the performance of this method with unsynchronized data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the editors, associate editors, and reviewers for their very helpful comments, which led to major improvements of the paper.

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-20-CE40-0025-01 (T-REX Project).

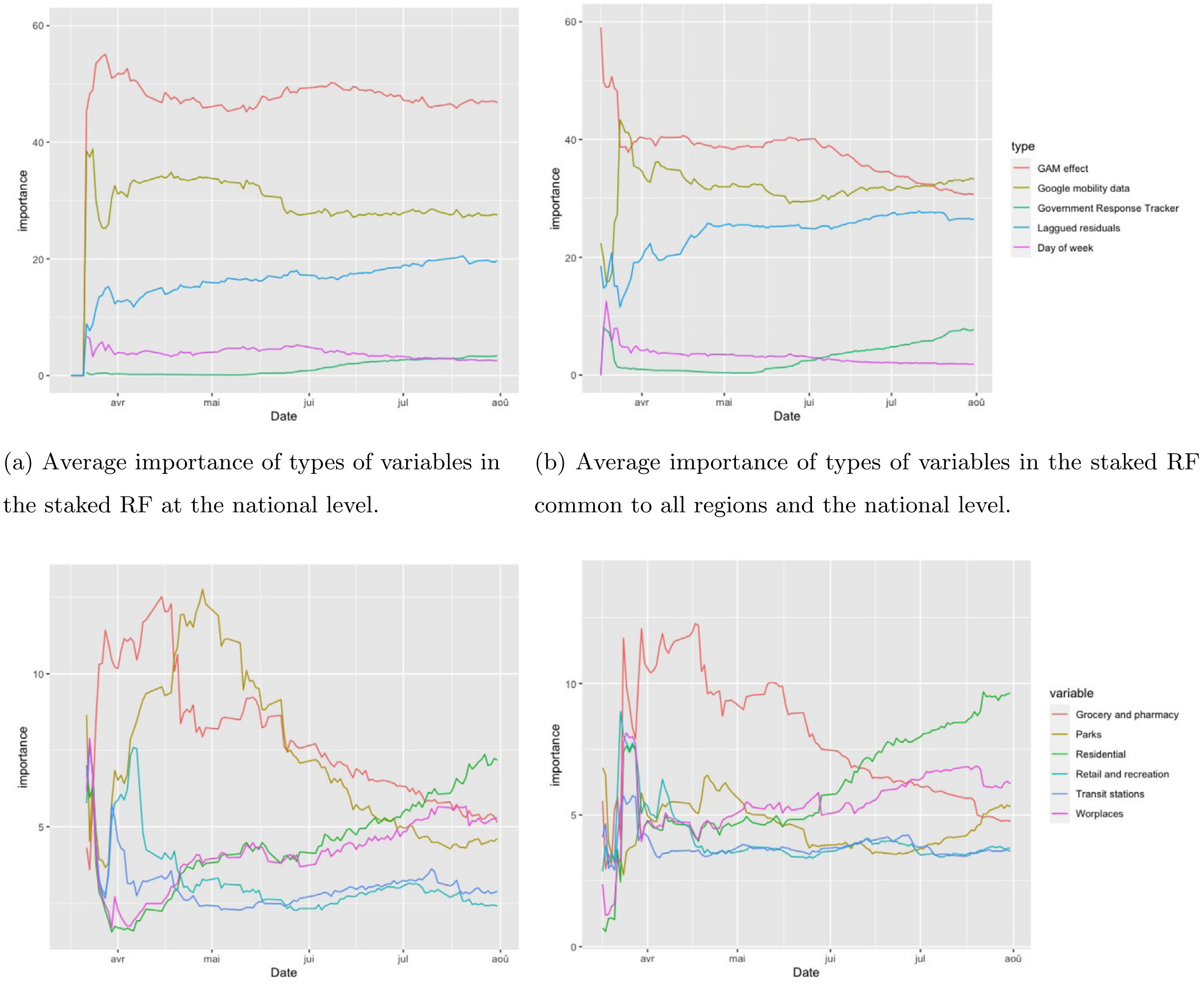
The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-19-CHIA-001 (Chair in Artificial Intelligence).

Appendix A. Variable selection for electricity load forecasting during the first COVID-19 lockdown

We allow the RFs to use many variables to make their predictions, including the usual calendar and weather variables, as well as mobility data, containment index, and estimated GAM effects, but without knowing a priori which will be relevant to predicting electricity consumption during the pandemic period. It is reasonable to assume that including all covariates might be detrimental to prediction given the high correlations between some variables and the small number of observations available for training the model, especially in the early days of the lockdown.

Variable selection for RFs. We allow the RFs to use many variables to make their predictions, including the usual calendar and weather variables, as well as mobility data, containment index, and estimated GAM effects, but without knowing a priori which will be relevant to predicting electricity consumption during the pandemic period. It is reasonable to assume that including all covariates might be detrimental to prediction given the high correlations between some variables and the small number of observations available to train the model, especially in the early days of lockdown. We want to exploit the fact that the number of observations increases rapidly by repeating the variable selection operation several times during the pandemic period in order to enrich the model if necessary. Moreover, we expect that the relevant variables might differ among regions, and thus we want to perform variable selection in a region by region manner.

In particular, we select the variables used to train the RFs for the following week's forecasts for a given region. Feature selection in RF is an ongoing field of research. State-of-the-art methods rely on the ranking of a variable importance measure, such as VSURF (Genuer, Poggi, & Tuleau-Malot, 2015), or on the permutation of variables, such as Boruta (Kursa & Rudnicki, 2010). These methods have high computational costs. In particular, VSURF is too slow to use in our context with numerous variable selection operations. We suggest an alternative approach for determining the relevant covariates by using techniques developed for variable selection in linear regression. Thus, we fit a linear model to predict the residuals of the GAM during the pandemic period using a LASSO penalty. Without prior knowledge of the necessary number of covariates to accurately forecast the electricity load, we design three models corresponding to different numbers of covariates. We fit a LASSO with



(a) Average importance of types of variables in the staked RF at the national level.

(b) Average importance of types of variables in the staked RF common to all regions and the national level.

(c) Average importance of mobility measures in the staked RF at the national level.

(d) Average importance of types of mobility measures in the staked RF common to all regions and the national level.

Fig. B.4. Evolution of the importance of the types of variables (top) and mobility measures (bottom) in the RFs trained on GAM residuals at the national level (left), and on GAM residuals for all regions and at the national level.

an ad-hoc penalty to select five and 10 covariates. This variable selection step is repeated every week. Finally, the predictions of the RFs using these five (10, respectively) covariates as inputs are combined with those by the RFs using all covariates as inputs using an expert aggregation method.

Before implementing this method for all regions and all quantiles, we evaluate its utility at predicting the 0.5 quantile of the national load using the available national data. We compare the MAPE values for the predictor obtained using Boruta, Lasso variable selection with an aggregation step, and the full model with 16 variables. The MAPE values for the lockdown and post-lockdown periods are presented in Table A.5.

The preliminary results indicate that variable selection only marginally affects the performance of the RFs, which highlights the robustness of the RFs against inputs with high dimensionality, even when trained on relatively small data sets. Therefore, we apply the RFs obtained using the full models with all covariates and estimated GAM effects as inputs.

Appendix B. Analysis of our method for electricity load forecasting during the first COVID-19 lockdown

B.1. Analysis of the stacked GAM and RF

Figs. B.4(a) and B.4(b) show the evolution of the average importance (for the different half-hours) of the variables for the stacked RF trained on residuals at the national level and on all residuals, respectively. The importance values of the different variables for a given model are normalized so their sum remains constant during the

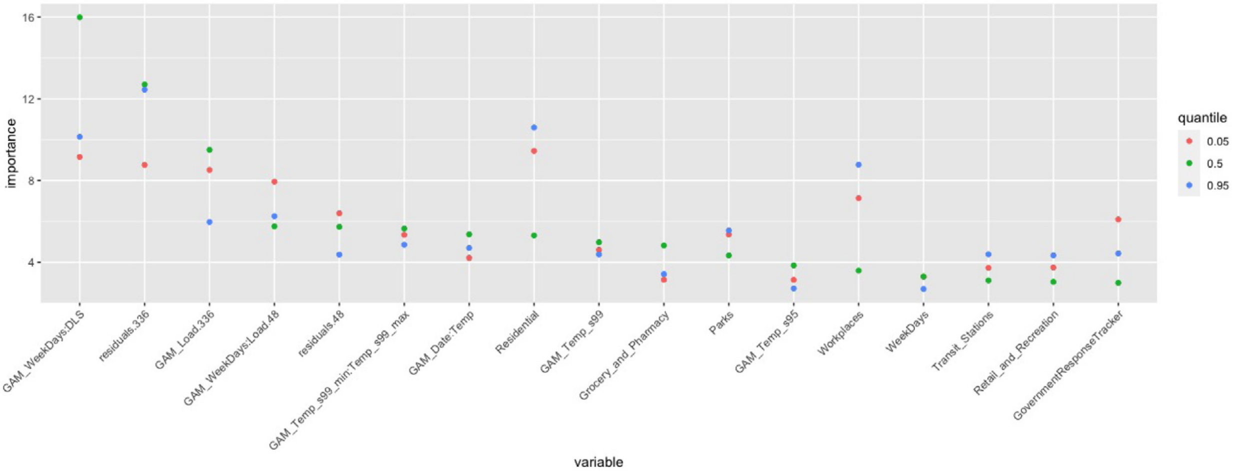
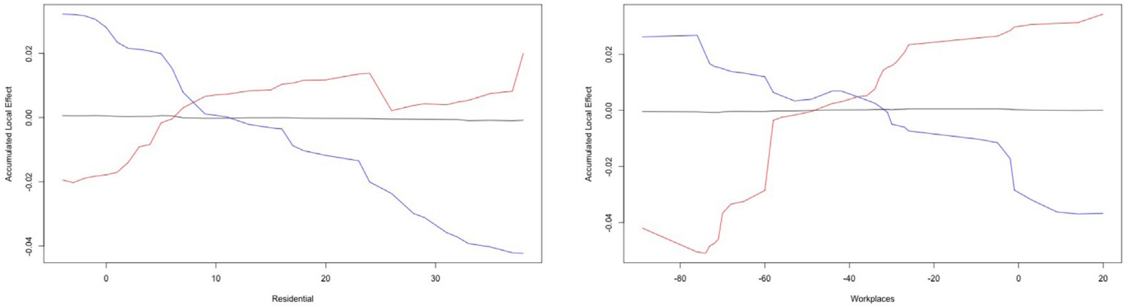


Fig. B.5. Importance of the variables in the stacked RFs for predicting quantiles 0.05, 0.5, and 0.95 of the GAM residuals at the national level. The variable "GAM_X" denotes the GAM effect corresponding to variable "X".



(a) Accumulated Local Effects of the relative frequentation of residential places.

(b) Accumulated Local Effects of the measure of relative frequentation of workplaces.

Fig. B.6. Accumulated local effects of the measure of relative frequentation of residential places (left) and workplaces (right) for the RF at the national level predicting the quantiles 0.05 (red), 0.5 (black), and 0.95 (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

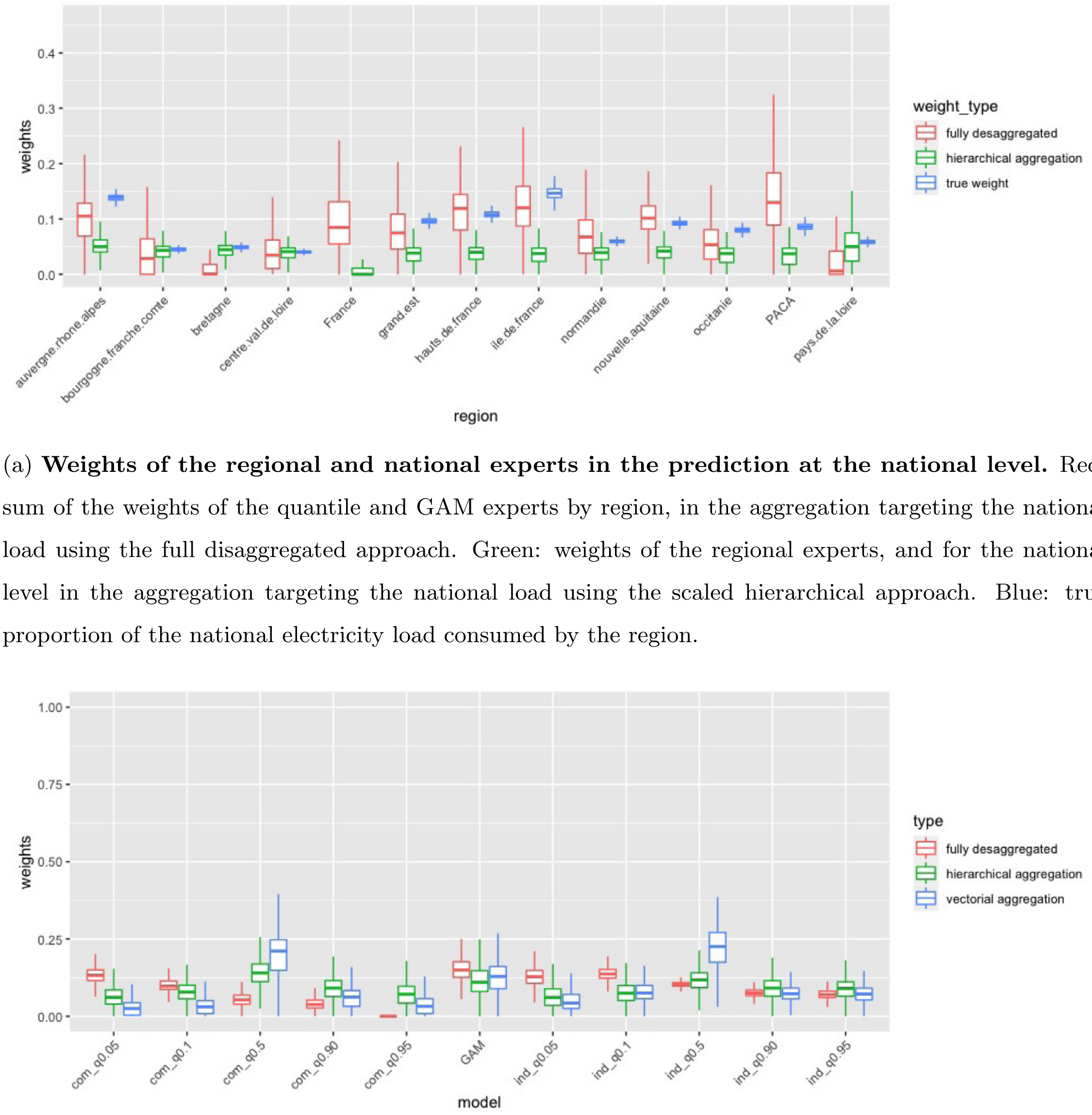
pandemic period, and equal to 100. In particular, if we denote $I_{v,t}$ and $I_{v,t}^{normalized}$ as the importance and normalized importance of variable v at time t, respectively, then at any time t, we have

$$I_{v,t}^{normalized} = 100 * \frac{I_{v,t}}{\sum_{variables\ v'} I_{v',t}}.$$

We group the variables into five categories: GAM effects, measures of mobility, government response tracker, lagged residuals, and day of the week. The importance of a group is simply the sum of the importance values of the variables in a group. The importance value for the mobility measures are shown in Figs. B.4(c) and B.4(d) (see Fig. B.4, Fig. B.7).

We note that the effects of the GAM are among the most important covariates for predicting the GAM residuals. Using these effects as covariates allows the transfer of information on the impacts of weather and calendar variables learned on the large data set of pre-pandemic observations. We observe a change in the importance of the different types of variables after the end of the lockdown, thereby indicating that the RFs can account for relative changes in the electricity consumption patterns. As time passes and the size of the training set for the RF increases, relevant variables such as the Government Response Tracker, relative occupation of residence, and grocery and pharmacies become more important for prediction. Conversely, spurious variables (e.g., the relative frequentation of parks, which is highly correlated with weather) are discarded as unimportant. Interestingly, the common RFs trained on residuals across all regions detect these relevant variables more quickly than the individual RFs trained solely on residuals at the national level. This highlights the benefit of multi-task learning in this sparse data context.

We also consider the relative importance of the variables in the stacked RFs for predicting the different quantiles. In particular, we consider the stacked RFs trained on the residuals at the national level for the pandemic period. We compute an importance measure for a given

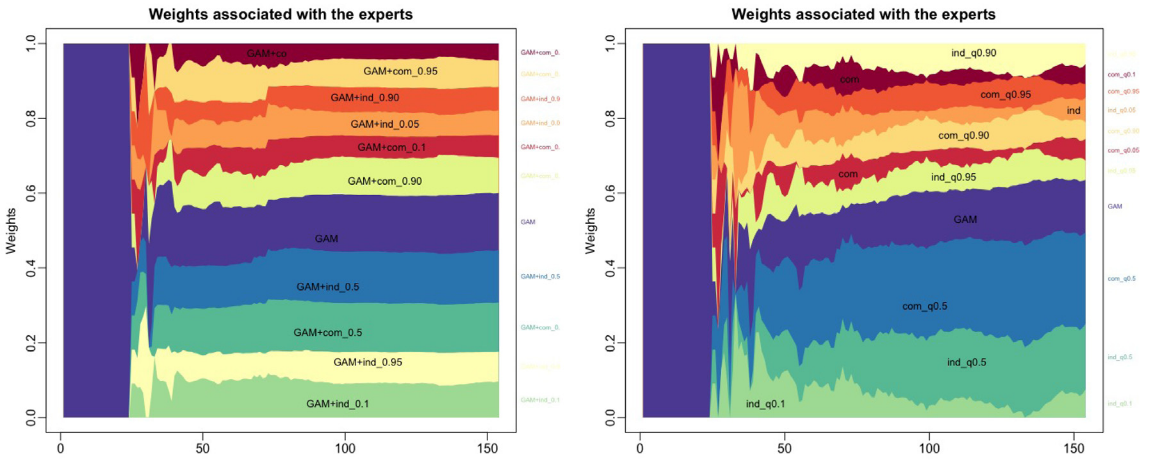


(a) **Weights of the regional and national experts in the prediction at the national level.** Red: sum of the weights of the quantile and GAM experts by region, in the aggregation targeting the national load using the full disaggregated approach. Green: weights of the regional experts, and for the national level in the aggregation targeting the national load using the scaled hierarchical approach. Blue: true proportion of the national electricity load consumed by the region.

(b) **Weights of the quantile and GAM experts in the prediction at the national level.** Weights of the quantile experts and the GAM expert in the aggregation targeting the national load using a full disaggregated approach (red), a hierachical aggregation approach (green), and a vectorial aggregation approach (blue).

Fig. B.7. Evolution of the weights of the quantile stacked GAM-RF experts and the GAM expert in the prediction of national load.

variable as the average increase in error in term of the pinball loss corresponding to a given quantile when the values of this variable are permuted at random (the error is computed over the training set). The importance values of the different variables are normalized so their sum is equal to 100. We compare the importance of the variables for predicting the 0.05, 0.5, and 0.95 quantiles in Fig. B.5. Variables that are important for predicting one quantile tend to be important for predicting the other quantiles. However, this is not the case for all variables. For example, the normalized loads for the relative frequentation of residential place and workplaces have the greatest importance for predicting the 0.05 and 0.95 quantiles. These variables have a very high (negative) correlation, where the frequentation of workplaces is very low during the lockdown period, and remains relatively low in the post-lockdown period during weekdays. By contrast, the frequentation of residential places is high during the



(a) Weights of the GAM and quantile experts in the first step of hierarchical aggregation, targeting the national load at 7:30 pm.

(b) Evolution of the weights of the GAM and quantile experts in the vectorial aggregation at 7:30 pm.

Fig. B.7. (*continued*).

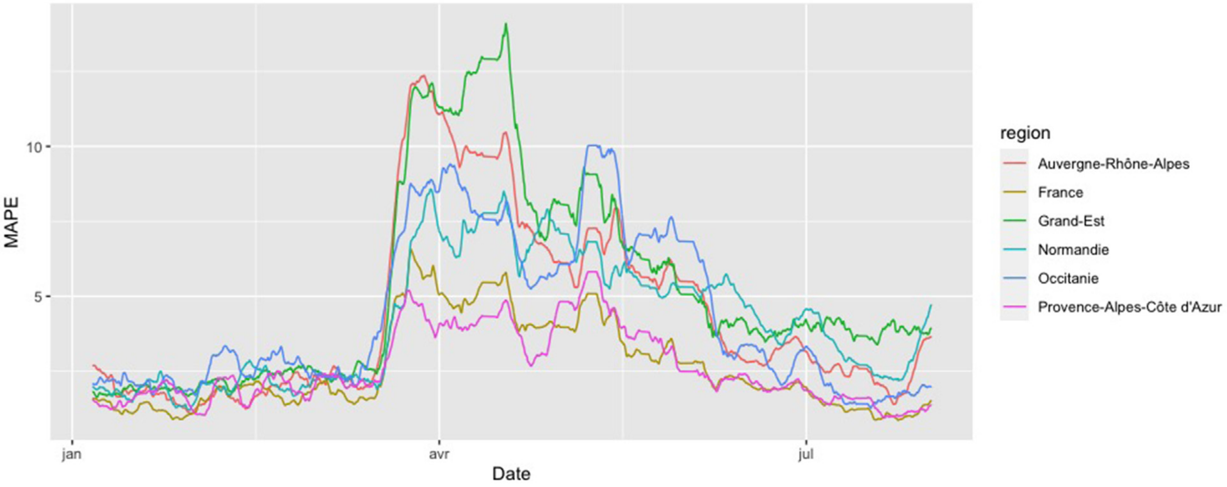


Fig. B.8. Weekly averaged MAPE of the normalized GAM for the regions Auvergne-Rhône-Alpes, Grand-Est, Normandie, Occitanie, and at the national level.

lockdown period and remains relatively high in the post-lockdown period during weekdays. The accumulated local effects of these variables shown in Fig. B.6 demonstrate that they have much larger impacts on the predictions of the two extreme quantiles than the median. However, we expect their effects to partially cancel each other out because of the correlations between these variables.

B.2. Analysis of online aggregation

Our results indicate that online aggregation is an efficient method for considering information available at a finer scale. It should be noted that the regional GAM has much larger average errors than GAM at the national level, as shown in Fig. B.8, which are due to the larger fluctuations present at the finer scale. Interestingly, aggregating these low-accuracy models obtains better performance than GAM at the national level, even in the pre-pandemic period, as shown in Table 4.

The fact that the scaled and unscaled hierarchical aggregation models obtain similar performance is somewhat counterintuitive given that in the scaled model, aggregation must learn the contributions of the different regions to the national consumption. According to the distribution of the weights for the regions in scaled hierarchical aggregation presented in Fig. B.7a, we note that the weights do not correspond to the proportion of electricity consumed by the regions (e.g., regions with low true weights, such as Provence-Alpes-Côte d'Azur, may receive more weight in aggregation than regions with high true weights, such as Île-de-France). Moreover, the weights in aggregation typically exhibit much more flexibility than the true weights, and this phenomenon is more striking in the fully disaggregated model. The high variability of the weights suggests that some of the models considered are fairly interchangeable. The fact that scaled hierarchical aggregation outperforms its unscaled counterpart both in the pre-pandemic and post-lockdown periods suggests that the flexibility provided by the second layer of aggregation used in the scaled model compensates for the lack of knowledge of the relative contributions of the different regions.

Fig. B.7b shows that all quantiles and GAM experts contribute to the predictions in both the fully disaggregated model and hierarchical aggregation model. By contrast, vectorial aggregation gives greater weights to GAM and the median stacked RF experts, which appear to be the most relevant experts across all regions. Fig. B.7 shows the weights given by aggregation for predicting the national load using only the national experts and the weights given by aggregation. Day 25 corresponds to the first day of the pandemic period and only the GAM forecast is available for the aggregation model before that day. We note that the weights in vectorial aggregation are highly unstable during the beginning of the lockdown. The performance of vectorial aggregation is worse during this period than all other aggregation models as well as the stacked RFs for predicting the median of residuals. This behavior reflects the fact that the impact of the pandemic differs strongly among regions, as shown in Fig. B.8. By contrast, vectorial aggregation performs best during the post-lockdown period and it appears to be a promising approach for predicting consumption under normal circumstances.

References

AECOM (2018). Energy demand research project: Early smart meter trials, 2007–2010. *Technical Report*, UK Data Service.

Amato, U., Antoniadis, A., De Feis, I., & Goude, Y. (2017). Estimation and group variable selection for additive partial linear models with wavelets and splines. *South African Statistical Journal*, *51*(2), 235–272.

Amato, U., Antoniadis, A., De Feis, I., Goude, Y., & Lagache, A. (2021). Forecasting high resolution electricity demand data with additive models including smooth and jagged components. *International Journal of Forecasting*, *37*(1), 171–185.

Anderer, M., & Li, F. (2022). Hierarchical forecasting with a top-down alignment of independent-level forecasts. *International Journal of Forecasting*, *38*(4), 1405–1414, Special Issue: M5 competition.

Balestriero, R., Pesenti, J., & LeCun, Y. (2021). Learning in high dimension always amounts to extrapolation. *Technical Report*, arXiv:2110.09485.

Brégère, M., & Huard, M. (2022). Online hierarchical forecasting for power consumption data. *International Journal of Forecasting*, *38*(1), 339–351.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Chapman & Hall/CRC.

Capezza, C., Palumbo, B., Goude, Y., Wood, S. N., & Fasiolo, M. (2021). Additive stacking for disaggregate electricity demand forecasting. *The Annals of Applied Statistics*, *15*(2), 727–746.

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. New York, NY, USA: Cambridge University Press.

De Vilmarest, J., & Goude, Y. (2022). State-space models for online post-covid electricity load forecasting competition. *IEEE Open Access Journal of Power and Energy*, *9*, 192–201.

Dong, Y., Zhang, H., Wang, C., & Zhou, X. (2021). Wind power forecasting based on stacking ensemble model, decomposition and intelligent optimization algorithm. *Neurocomputing*, *462*, 169–184.

Fan, S., & Hyndman, R. J. (2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, *27*(1), 134–141.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Farrokhabadi, M., Browell, J., Wang, Y., Makonin, S., Su, W., & Zareipour, H. (2022). Day-ahead electricity demand forecasting competition: post-covid paradigm. *IEEE Open Access Journal of Power and Energy*, *9*, 185–191.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., & Goude, Y. (2021). Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, *116*(535), 1402–1412.

Gaillard, P., & Goude, Y. (2016). Opera: Online prediction by expert aggregation. URL: https://CRAN.R-project.org/package=opera, R Package Version, 1.

Gaillard, P., Stoltz, G., & Van Erven, T. (2014). A second-order bound with excess losses. In *Conference on learning theory* (pp. 176–196).

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2015). VSURF: An r package for variable selection using random forests. *The R Journal*, *7*(2), 19–33.

Gilbert, C., Browell, J., & Stephen, B. (2023). Probabilistic load forecasting for the low voltage network: Forecast fusion and daily peaks. *Sustainable Energy, Grids and Networks*, *34*, Article 100998.

Goehry, B., Goude, Y., Massart, P., & Poggi, J.-M. (2019). Aggregation of multi-scale experts for bottom-up load forecasting. *IEEE Transactions on Smart Grid*, *11*(3), 1895–1904.

Goude, Y., Nedellec, R., & Kong, N. (2013). Local short and middle term electricity load forecasting with semi-parametric additive models. *IEEE Transactions on Smart Grid*, *5*(1), 440–446.

Khairalla, M. A., Ning, X., Al-Jallad, N. T., & El-Faroug, M. O. (2018). Short-term forecasting for energy consumption through stacking heterogeneous ensemble learning model. *Energies*, *11*(6), 1605.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*(11), 1–13.

Laptev, N., Yu, J., & Rajagopal, R. (2018). Applied timeseries transfer learning.

Meinshausen, N., & Ridgeway, G. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(6).

Moon, J., Jung, S., Rew, J., Rho, S., & Hwang, E. (2020). Combination of short-term load forecasting models based on a stacking ensemble approach. *Energy and Buildings*, *216*, Article 109921.

Obst, D., Ghattas, B., Claudel, S., Cugliari, J., Goude, Y., & Oppenheim, G. (2022). Improved linear regression prediction by transfer learning. *Comput. Stat. Data Anal.*, *174*(C).

Obst, D., de Vilmarest, J., & Goude, Y. (2021). Adaptive methods for short-term electricity load forecasting during COVID-19 lockdown in France. *IEEE Transactions on Power Systems*, *36*(5), 4754–4763.

Olivas, E. S., Guerrero, J. D. M., Sober, M. M., Benedito, J. R. M., & Lopez, A. J. S. (2009). *Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques - 2 Volumes*. Hershey, PA: Information Science Reference - Imprint of IGI Publishing.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Schellong, W. (2011). *Energy demand analysis and forecast* (pp. 101–120). BoD–Books on Demand.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75–85.

Wang, Y., Chen, Q., Hong, T., & Kang, C. (2019). Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid*, *10*(3), 3125–3148.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Wood, S. (2017). *Generalized additive models: an introduction with R* (2nd ed.). Chapman and Hall/CRC.

Xenochristou, M., & Kapelan, Z. (2020). An ensemble stacked model with bias correction for improved water demand forecasting. *Urban Water Journal*, *17*(3), 212–223.

Zhai, B., & Chen, J. (2018). Development of a stacked ensemble model for forecasting and analyzing daily average PM2. 5 concentrations in Beijing, China. *Science of the Total Environment*, *635*, 644–658.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., et al. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76.

Ziel, F. (2022). Smoothed bernstein online aggregation for short-term load forecasting in ieee dataport competition on day-ahead electricity demand forecasting: post-covid paradigm. *IEEE Open Access Journal of Power and Energy*, *9*, 202–212.