Aggregation of nonlinearly enhanced experts with application to electricity load forecasting

A. Incremona[a,*], G. De Nicolao[a], F. Fusco[b], B. J. Eck[b], S. Tirupathi[b]

[a] *Department of Industrial and Information Engineering, University of Pavia, Via Adolfo Ferrata 5, 27100, Pavia, Italy*
[b] *IBM Research Europe, Ireland*

Please cite this as: A. Incremona, G. De Nicolao, F. Fusco, B.J. Eck, S. Tirupathi, Aggregation of nonlinearly enhanced experts with application to electricity load forecasting, Applied Soft Computing, 2021, 107857, https://doi.org/10.1016/j.asoc.2021.107857.

Abstract

Combining the predictions of different base experts is a well known approach used to improve the accuracy of time series forecasts. Forecast aggregation is becoming crucial in many fields, including electricity forecasting, as Internet of Things and Cloud technology give access to larger numbers of sensor data, time series and predictions from external providers. In this context, it is not uncommon that the failure of some experts causes relevant drops in the performances of the aggregated forecast when classical techniques based on linear averaging are applied. This might be a symptom of suboptimality of the individual experts, that do not fully exploit important predictors, e.g. calendar features that play a major role in the electrical demand profiles. In this work, we therefore present two non-linear strategies to obtain aggregated forecasts, starting from the availability of a set of base experts and the knowledge of some relevant predictor variables. The first approach, called aggregation of enhanced experts (AEE), enhances each individual expert and then feeds the enhanced forecasts into classical linear aggregation techniques. In the second approach, called enhanced aggregation of experts (EAE), the expert forecasts are nonlinearly combined with the predictor variables through an Artificial Neural Network (ANN). The case of missing expert forecasts is also considered via a statistically-based imputation method. The short-term prediction of German

*Corresponding author
Email address: alessandro.incremona01@universitadipavia.it (A. Incremona)

electrical load is used as a case study. Twelve base experts are enhanced with respect to calendar features, i.e. daytime and weekday. Compared to state-of-the-art aggregation methods applied to the not-enhanced set of experts, the proposed approaches not only improve the accuracy of aggregated forecast (up to 25% reduction of MAPE and RMSE), but are also robust with respect to missing experts.

Keywords: load demand, neural networks, missing features, forecast aggregation, enhancement

1. Introduction

The easiness of access to big volumes of information, the improved communication technologies and the increasing automation of the processes, besides representing remarkable advantages, pose also new challenges for what concerns complexity and effective management of data. Among the possible data streams, it will become more and more frequent the availability of multiple predictions originated by external experts, a notable example being offered by the energy market where load forecasts at local/global level and short/medium/long term are produced by multiple subjects [1]. Similar scenarios are expected to grow in many other contexts in view of the widespread adoption of "Internet-of-Things" (IoT) technologies relying on the physical and virtual interconnection between sensing devices that gives access to a large number of data useful for decision making in several industries (e.g. smart grids, transportation systems, building management).

In the energy field, accurate predictions of the power demand are needed in order to ensure the real time balance between production and consumption, which is necessary for stable electricity supply, system security, efficient allocation of power generation. Herein, we consider the problem of obtaining a reliable short term load forecast, based on the availability of several ready-to-use forecasts from external providers, herein called 'base experts' [2, 3]. In view of their ready-to-use nature and the impossibility of retraining them, the expert

forecasts are called 'secondary data'. By contrast, the features (predictors) associated with the target to be forecast are called 'primary data'. It is to be noted that in the case of time series, certain kinds of primary data are always available, since some features, such as daytime, weekday, and day of the year are typically known. This means that, in alternative to the simple aggregation of the secondary data, taken on an 'as is' basis, a more ambitious strategy can be pursued, i.e. the enhancement of the base forecasts by judicious use of the primary data.

In the forecasting community, the combination of multiple learning algorithms and models is commonly adopted in order to improve forecasts [4]. Ensemble learning methods such as bagging, boosting and stacking rely on training several base learners in such a way that the combination of their outputs leads to a reduction of the variance (bagging) or bias (boosting): see [5] for a comprehensive review on ensemble methods in machine learning, [6] for applications of bagging methods on real-world and simulated data, [7] for a comparison between bagging and boosting approaches, [8] for a detailed description of the advantages of stacking ensemble strategies and [9] for applications of ensemble learning methods in the field of load forecasting. However, when predictions come from external sources, it may be not possible to access their generating models and even information on their structure could be missing. Nonetheless, improvements could still be obtained by suitable aggregation methods. In the literature, a variety of techniques have been explored, ranging from average-based ones to more sophisticated machine learning approaches: in [10] a variety of algorithms for combining sister forecasts are proposed and compared, in [11] the problem of aggregating predictions is considered in a probabilistic scenario, in [12] forecast aggregation is applied in the context of hierarchical modeling while [13] compares weighting-based aggregation schemes to traditional ones. These techniques have proven to be successful in competitions [14] and are applied to different fields, such as finance [15], weather [16] and load demand forecasts [17].

However, there are issues that, though commonly encountered in practice,

have not yet been satisfactorily addressed. The available predictions may come from models trained on old data, with suboptimal choice or use of some features. If the forecast is externally generated it may be impossible to precisely diagnose these shortcomings and even more impossible to fix them. Nevertheless, one may not want to drop the expert altogether, given that its generating model could exploit some relevant features that would be otherwise unavailable. Under these biases, elementary aggregation methods help to reduce the variance of the individual predictions, but the potentialities of each expert are fatally underutilized.

The main novelty of this work with respect to the traditional forecast aggregation literature is the introduction of two 'enhancement strategies' so as to fix the experts' shortcomings and improve the final aggregated prediction. In particular, given a benchmark problem, a preliminary exploratory analysis is performed in order to assess the expert shortcomings. Then, an aggregation scheme that uses the primary data in order to enhance the secondary ones, i.e. the expert forecasts, is introduced, so as to fix their weaknesses and achieve a better final aggregated prediction. Finally, the aggregated predictions are compared to state-of-the-art aggregation methods applied to the plain, i.e. not-enhanced, experts.

The learning framework assumes the availability of a training dataset made of primary and secondary data. The primary data include target and feature variables, while the secondary ones gather predictions made by external experts. It is worth remarking that the features available in the primary data may cover only partially those used by the experts to produce their predictions. For this reason, it is not convenient to drop the experts and retrain a model based on the primary data alone, because valuable information incorporated in the experts may go lost.

Two approaches are considered for blending the experts with the primary variables. The first one, called aggregation of enhanced experts (AEE), is a two-step approach. First, each expert is individually enhanced, accounting for the primary data. In the second step, the enhanced experts are aggregated by

weighted average techniques. The second approach, conversely, consists of the enhanced aggregation of experts (EAE) through an Artificial Neural Network trained using primary and secondary data in order to predict the target. The possible occurrence of missing experts is addressed by a statistical imputation technique.

The short-term prediction of German electrical load is used as a case study. In this context, the secondary variables to be aggregated are the predictions of twelve experts, whose underlying models are supposed to be unknown, while the primary variables are the calendar features, i.e. daytime (quarter-hourly measured) and weekday. The comparison of the two enhanced aggregation approaches to state-of-art aggregation methods shows that enhancement brings a significant improvement.

The paper is organized as follows: in Section 2, the benchmark dataset is described. In Section 3 the *Expert enhancement/aggregation problem* is formulated and the two enhancement/aggregation strategies are described. In Section 4, the different methods are applied to the benchmark data in both 'full information' and 'missing-experts' scenario and the results are shown. In Section 5, the main conclusions are summarized and discussed.

2. Benchmark problem: German electricity load forecasting

Let consider the forecasting of the German electric load demand, denoted by

$$y(t_k) \in \mathbb{R}, \quad t_k - t_{k-1} = 0.25 \text{ hour}, \forall k \tag{1}$$

For the sake of simplicity, hereafter the shorthand notation $y(k)$ will be used in place of $y(t_k)$.

The actual load demand, recorded from January 1 to September 6, 2019, is displayed in Fig. 1. The quarter-hourly data (MW) were downloaded from the Entso-E transparency platform [18].

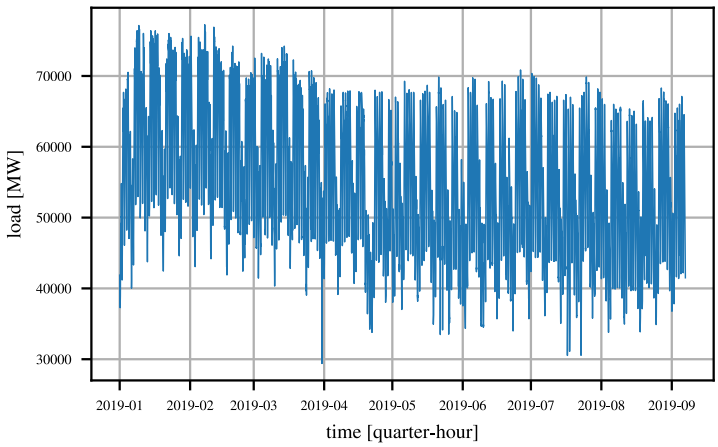

Figure 1: German quarter-hourly electric load demand from January 1^{st} to September 6^{th}, 2019.

During this period, the forecasts provided by twelve experts, trained over the years 2017 and 2018, are available:

$$\xi_i(k) \in \mathbb{R}, \quad i \in \mathcal{M}, \quad \mathcal{M} = \{1, \ldots, 12\} \tag{2}$$

All these experts are underpinned by models belonging to the Generalized Additive Models (GAMs) category [19], but they differ either in the choice of features or calibration procedures. In the literature, the term 'sister forecasts' has been used to denote families of forecasters sharing a common structure [10]. The features considered by each expert are summarized in Table 1. Among them, there is also the binary d_{type} feature that identifies special days such as holidays. For simplicity of notation, from now on we will refer to the 'experts forecasts' just as 'experts'.

The model structures underpinning the experts are supposed to be unknown which means that their possible weaknesses are not known *a priori* but can nevertheless be assessed *ex post* from the observed data. A preliminary exploration can be performed by inspecting the prediction residuals $r_i := y - \xi_i$. In particular, the joint bivariate distributions of the residuals for all possible pairs of experts can be visualized through the scatter matrix displayed in Fig. 2, where

also the correlation coefficients are reported. There are pairs of highly correlated residuals (e.g. experts 2 and 3) and other pairs whose correlation is poor (e.g. experts 3 and 11). Of course, if all expert residuals were highly correlated, there would be no room for improvement by aggregation.

Figure 2: Forecasting residuals of the 12 experts and observed German loads: scatter matrix and correlation coefficients.

The correlations between residuals and observations are shown in red in the last column of Fig. 2 and the correlation coefficients are shown in the last row. A highly performing expert will be characterized by low correlation and small residual variance, see e.g. experts 9 and 11, whose scatter plots appear horizontal and thin. Conversely, a significant correlation, not only is a symptom

of suboptimality, but could be leveraged to reduce the error, a notable example being represented by expert 12. The twelve univariate residual distributions can be compared through their boxplots, see Fig. 3. It appears that in all cases the residuals are negatively biased, meaning that in the considered period all the experts tend to overestimate the load.

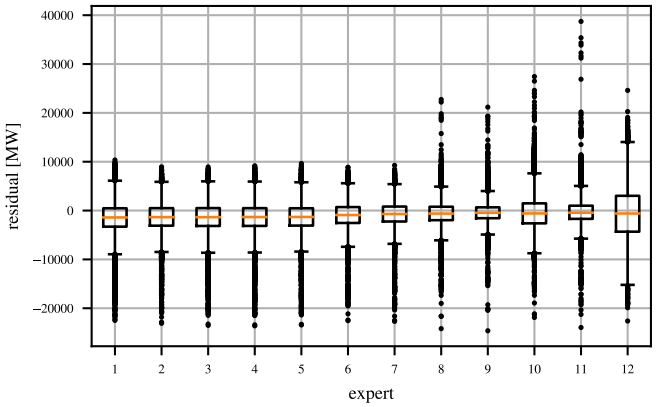

Figure 3: Boxplots of the forecasting residuals of the 12 experts.

A visual example of the weaknesses of some of the experts is given in Fig. 4, where experts 1 and 10 are plotted over a sample week together with the observations. Both experts fail to catch the night trough of the load demand, expert 10 overestimates the morning demand during the weekend and expert 1 exhibits a daily pattern which is way too smooth with respect to the actual shape of the load demand. A more systematic insight can be gained by looking at Fig. 5, reporting the scatter plots of the quarter-hour of the day against the residuals of experts 1 and 10 on Mondays and Saturdays. The consistent patterns seen in these scatter plots are not specific to these two experts, but analogous patterns, depending on the day of the week, are found also in the residuals of the other experts.

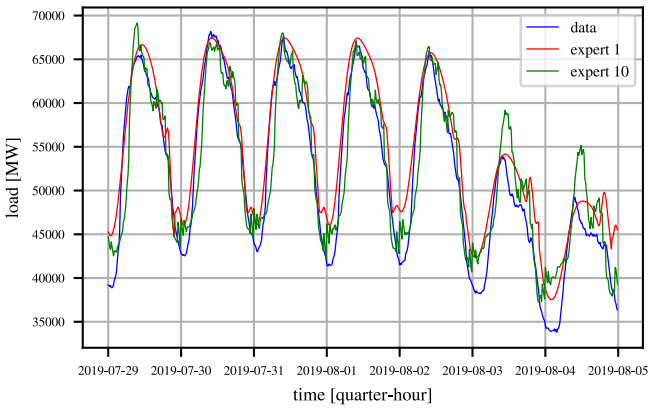

Figure 4: Observed German loads (blue), Expert 1 forecasts (red), and Expert 10 forecasts (green) during a one week period.

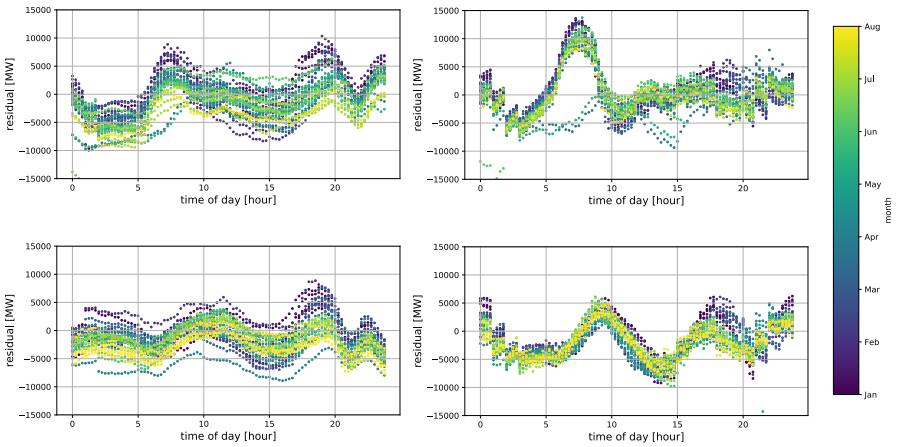

Figure 5: Scatter plots of forecasting residuals of Expert 1 (left panels) and 10 (right panels) against time of day on Monday (top panels) and Saturday (bottom panels).

Table 1: Expert models' features. T and I represent respectively the temperature and the normal solar irradiance, t_{year} and t_{day} are calendar features associated with the time of year and the time of day, and d_{type} is the binary feature identifying special days such as holidays.

Expert	**Features**
Expert 1	$t_{year},\ t_{day},\ T(k),\ \xi(k-24),\ d_{type}$
Expert 2	$t_{year},\ t_{day},\ T(k),\ \xi(k-24),\ d_{type}$
Expert 3	$t_{year},\ t_{day},\ T(k),\ \xi(k-24),\ I(k),\ d_{type}$
Expert 4	$t_{year},\ t_{day},\ I(k),\ T(k-8),\ T(k-16),$ $T(k-24),\ \xi(k-8),\ \xi(k-16),\ \xi(k-24),\ d_{type}$
Expert 5	$t_{year},\ t_{day},\ I(k),\ T(k-8),\ T(k-16),$ $T(k-24),\ \xi(k-8),\ \xi(k-16),\ \xi(k-24),\ d_{type}$
Expert 6	$t_{year},\ t_{day},\ I(k),\ T(k),\ T(k-1),\ T(k-2),$ $T(k-3),\ T(k-4),\ T(k-24),\ \xi(k-24),\ d_{type}$
Expert 7	$t_{year},\ t_{day},\ I(k),\ T(k-24),$ $T(k-48),\ \xi(k-24),\ \xi(k-48),\ d_{type}$
Expert 8	$t_{year},\ t_{day},\ I(k),\ T(k),\ T(k-8),\ T(k-16),$ $T(k-24),\ \xi(k-12),\ \xi(k-16),\ \xi(k-24),\ d_{type}$
Expert 9	$t_{year},\ t_{day},\ I(k),\ T(k),\ T(k-4),\ T(k-8),$ $T(k-12),\ T(k-16),\ T(k-20),$ $T(k-24),\ \xi(k-8),\ \xi(k-16),\ \xi(k-24),\ d_{type}$
Expert 10	$T(k),\ T(k-4),\ T(k-8),\ T(k-12),\ T(k-16),$ $T(k-20),\ T(k-24),\ \xi(k-12),\ \xi(k-24),\ d_{type}$
Expert 11	$I(k),\ T(k),\ T(k-4),\ T(k-8),\ T(k-12),$ $T(k-16),\ T(k-20),\ T(k-24),\ \xi(k-8),$ $\xi(k-8),\ \xi(k-12),\ \xi(k-24),\ d_{type}$
Expert 12	$T(k-24),\ T(k-48),\ T(k-72),$ $\xi(k-24),\ \xi(k-48),\ \xi(k-72),\ d_{type}$

From this exploratory analysis it emerges that there exists room for improvement, provided that the the weaknesses of the experts are fixed by a suitable enhancement technique. These weaknesses are not necessarily due to inadequate design of the experts. In fact, the distribution of the observations could have changed over time. In particular, the correlation between some features and the target might have changed.

The exploratory analysis highlighted also a possible way to achieve an improvement, since the residuals are correlated with calendar features such as the weekday and the quarter-hour of the day, that can be regarded as exogenous features, usable to enhance the base experts.

3. Expert enhancement and aggregation

3.1. Problem statement

Motivated by the German load forecasting example, we are now in a position to state the *Expert enhancement/aggregation problem*. In the following, $\mathcal{Tr}$ and $\mathcal{Te}$ will denote the set of time indices associated to the training and test set, respectively.

Problem Statement. *Given the training data* $\{y(k), \boldsymbol{\xi}(k), \boldsymbol{x}(k)\}$, $k \in \mathcal{Tr}$, *where* $\boldsymbol{\xi}(k) = [\xi_1(k) \dots \xi_m(k)]^T$ *is the vector of the experts and* $\boldsymbol{x}(k) = [x_1(k) \dots x_p(k)]^T$ *are exogenous features, devise an* expert enhancer $H(\cdot)$ *that provides an estimate* $\hat{y} = H(\boldsymbol{\xi}, \boldsymbol{x})$ *of the target variable* y.

In the German load problem, this implies to develop a forecaster $\hat{y}$ blending the contributions of the twelve experts ξ_i, $i = 1, \dots, 12$, with some exogenous features x_i, $i = 1, 2$, that, in the case considered, are chosen as the quarter-hour of day and the weekday.

The training and the testing steps require suitable datasets. The whole available dataset does not cover an entire year and using consecutive periods of time for the training and testing phases could lead to misleading results because of the yearly seasonality. This motivated the adoption of a random selection of training days and testing days. The quarter-hourly load data were organized in 96-sample blocks, each of one corresponding to a single day. Then, one third of these blocks, randomly selected, was reserved for testing purposes. Note that, in its essence, our learning problems will be a static one. Indeed, the enhancement of the experts does not involve data other than those at that specific day, e.g. lagged observations from previous days. The only information

used to modify the experts' forecasts are the weekday and the daytime of the forecasts themselves. Of course, our approach is well suited to correct experts' biases that are not evolving rapidly, in which case adaptive methods would have to be elaborated and validated on datasets much longer than the one considered in this paper.

In the previous arrangement, the one-day 96-sample blocks are randomly assigned to training or testing, irrespective of causality constraints. In order to assess performances also when a strictly causal training-testing splitting is adopted, we considered a second scheme in which the algorithms are trained on data ranging from January 1 to July 31 and tested on the remaining load data. As observed above, this is a kind of stress test because the training data do not cover an entire year so that some robustness against seasonal drifts is called upon.

Typical performance metrics are the Mean Absolute Percentage Error

$$MAPE = \frac{100}{n} \sum_{k \in \mathcal{T}e} \left| \frac{\hat{y}(k) - y(k)}{y(k)} \right| \tag{3}$$

and the Root Mean Square Error

$$RMSE = \sqrt{\frac{\sum_{k \in \mathcal{T}e} \left(\hat{y}(k) - y(k)\right)^2}{n}} \tag{4}$$

In the following, we propose two approaches for the *Expert enhancement/aggregation problem*: the aggregation of enhanced experts (AEE) and the enhanced aggregation of experts (EAE).

3.2. Aggregation of enhanced experts (AEE)

Individual enhancers

This approach splits the enhanced aggregation process in two phases. First, each expert is fed into an individual enhancer $h_i(\cdot, \cdot)$ that, exploiting the exogenous features, yields the enhanced experts

$$\hat{\xi}_i = h_i(\xi_i, \boldsymbol{x}) \tag{5}$$

Then, the enhanced experts undergo a final aggregation step that yields the final prediction

$$\hat{y} = G(\hat{\xi}) \tag{6}$$

where $G(\cdot)$ is the aggregator function and $\hat{\boldsymbol{\xi}} = [\hat{\xi}_1 \dots \hat{\xi}_m]$ is the vector of the enhanced experts. Different strategies can be adopted for the design of the individual enhancers. Since the basic idea is a sort of 'retuning' of the expert, based on few exogenous features, the enhancer should privilege simplicity and robustness. The simplest enhancers are the additive, multiplicative and affine ones. Since they implement basic correction schemes they would be the first to be included in the 'enhancers toolbox' used to compute the enhanced experts that will undergo the second step, i.e. aggregation. In many contexts, one could take advantage of more complex enhancers, able to model arbitrary nonlinear correction schemes. Two examples are the regularization-based tensor spline and the MLP Artificial Neural Network. Of course other nonlinear models could be used to enrich the suite of possible enhancers. Below, five enhancers are considered: additive, multiplicative, affine, tensor spline and MLP ANN.

Additive Enhancer. A simple enhancement consists of an additive correction term β_i that depends nonlinearly on the exogenous features:

$$\hat{\xi}_i = \xi_i + \beta_i(\boldsymbol{x}), \quad i \in \mathcal{M} \tag{7}$$

For the benchmark problem,

$$\hat{\xi}_i = \xi_i + \beta_i(t_{day}, d_{week}), \quad i = 1, ..., 12 \tag{8}$$

where the weekday $d_{week} \in \{\text{Monday}, \dots, \text{Sunday}\}$ is a categorical variable and the quarter-hour time t_{day}, $1 \leq t_{day} \leq 96$, is treated as a real variable on the circle. For this reason, training $\beta(\cdot, \cdot)$ implies to train seven periodic longitudinal functions $\beta_i(\cdot, d_{week})$, $d_{week} \in \{\text{Monday}, \dots, \text{Sunday}\}$. In order to account for the 24-hour periodicity, cyclic penalized cubic B-splines (cyclic P-splines) were adopted as basis functions (see Appendix). The squared second derivative

was used as regularization penalty. With this choice, the regularization hyperparameter determines the rate of change of the correction term throughout the 24 hours.

Multiplicative Enhancer. Another type of enhancement consists of a multiplicative correction term α_i that depends nonlinearly on the exogenous features:

$$\hat{\xi}_i = \xi_i \alpha_i(\boldsymbol{x}) \tag{9}$$

For the benchmark, this translates into the model

$$\hat{\xi}_i = \xi_i \alpha_i(t_{day}, d_{week}) \quad i = 1, ..., 12 \tag{10}$$

The structure and training of α_i is analogous to that of β_i in the additive enhancer.

Affine Enhancer. When both additive and multiplicative correction terms are employed, we obtain the affine enhancer:

$$\hat{\xi}_i = \xi_i + \beta_i(\boldsymbol{x}) + \xi_i \alpha_i(\boldsymbol{x}) \tag{11}$$

It is worth observing that the enhancer has a GAM structure, meaning that a backfitting iteration [19] [20] can be used for its training: at each step, only the estimation of either α_i and β_i is performed. For the benchmark, the affine enhancer takes the form

$$\hat{\xi}_i = \xi_i + \beta_i(t_{day}, d_{week}) + \xi_i \alpha_i(t_{day}, d_{week}) \tag{12}$$

The structures of α_i and β_i remain the same as in the additive and multiplicative enhancers. By resorting to backfitting, at each step only the estimation of longitudinal functions of the quarter-hour time is required.

Tensor Spline Enhancer. A more flexible enhancer is obtained by resorting to a generic nonlinear function described by a tensor spline model:

$$\hat{\xi}_i = \xi_i + f_i^{\text{tensor}}(\xi_i, \boldsymbol{x}) \tag{13}$$

For the benchmark,

$$\hat{\xi}_i = \xi_i + f_i^{\text{tensor}}(\xi_i, t_{day}, d_{week}) \tag{14}$$

where $f_i^{tensor}(\cdot, \cdot, d_{week})$, $d_{week} \in \{\text{Monday}, \ldots, \text{Sunday}\}$ are seven cubic bivariate tensor splines, describing a function of two variables (expert and the quarter-hourly time of day). The tensor functions are built using P-splines for the expert component direction and cyclic P-splines for the quarter-hour time of day.

MLP-ANN Enhancer. Another flexible enhancer uses as correction term a Multi-Layer Perceptron Artificial Neural Network (MLP-ANN) f^{MLP} with input vector $[\xi_i \quad \boldsymbol{x}^T]^T$:

$$\hat{\xi}_i = \xi_i + f_i^{\text{MLP}}([\xi_i \quad \boldsymbol{x}^T]^T) \tag{15}$$

For the benchmark, we let $x = t_{day}$ and seven ANN's are trained, one for each weekday.

Hyperparameters calibration

The first four enhancers belong to the Generalized Additive Models (GAMs) family and were implemented using the `Python`'s library `pyGAM`. The Artificial Neural Network was implemented using the `Python`'s library `scikit-learn`.

For what concerns the splines terms of the GAMs-based enhancements, a fairly large number of basis functions was used, entrusting the regularization task to the penalty hyperparameter. In particular, 12 cyclic P-splines were used for the daily periodicity (one spline every two hours) and 20 P-splines for the expert prediction variable. The regularization parameters were tuned, for each expert, using Generalized Cross Validation (GCV).

The MLP-ANN enhancer hyperparameters, consisting of the number of hidden layers, the number of neurons for each layer, the activation function and the L2-regularization term α, were tuned through 5-fold cross-validation, leading to an architecture with one hidden layer with 12 neurons, hyperbolic tangent activation function and $\alpha = 0.01$ for each expert. The optimization was carried out by the `adam` stochastic gradient-based optimizer proposed by [21].

Experts aggregation

The second and final step of the AEE scheme is the aggregation of the enhanced predictors. In the following, five possible aggregation methods are reviewed.

Simple average. Despite its simplicity, the arithmetic mean is a widely used method for aggregating individual experts. For some economic time series it was even found that only few aggregation schemes outperformed the simple equally weighted average of expert forecast [22]. The simple average is a robust method and deals seamlessly with the problem of missing experts.

Winsorized average. The Winsorized average (WA) is a more robust version of the simple average method, where the two extreme experts are replaced by the second largest and the second smallest experts.

Trimmed average. The Trimmed average (TA) is another robust extension of the simple average method, where at each time the two extreme experts are discarded from the computation.

Constrained Least Squares. The Constrained Least Squares (CLS) average is a weighted linear combination, whose weights are chosen so as to minimize the sum of squared residuals, constraining the weights to be nonnegative and to sum up to one. More precisely,

$$\hat{y} = \boldsymbol{\theta}^{\boldsymbol{T}}\hat{\boldsymbol{\xi}}, \quad \hat{\boldsymbol{\xi}} = \left[\hat{\xi}_1 \dots \hat{\xi}_m\right]^T \tag{16}$$

$$\theta = \arg\min_{\theta} \sum_{k \in \mathcal{T}r} \left(y(k) - \boldsymbol{\theta}^T \hat{\xi}(k)\right)^2 \tag{17}$$

$$\text{s.t.} \quad \theta_i \geq 0, \quad \forall i \tag{18}$$

$$\sum_{i \in \mathcal{M}} \theta_i = 1 \tag{19}$$

This not only has the advantage of preventing weights instability that may ensue from the the collinearity of the experts, but guarantees a more interpretable

model as well [10], since it allows to detect the most important experts by giving them higher coefficients, while silencing uninformative predictors by assigning them coefficients close to zero. In this way, the algorithm itself selects the best experts among all. For the benchmark problem, the optimal weights were computed through quadratic programming, using the `Python`'s library `cvxopt`.

Multi-layer Perceptron. In order to assess the potential of nonlinear aggregation, a one-hidden layer MLP-ANN aggregation method is considered as well:

$$\hat{y} = f^{\mathrm{MLP}}(\hat{\boldsymbol{\xi}}) \tag{20}$$

The structure of this MLP-ANN model was calibrated through 5-fold cross-validation and consists of one hidden layer with 12 neurons, hyperbolic tangent activation function and regularization penalty $\alpha = 1$.

3.3. Enhanced Aggregation of Experts (EAE)

The second approach, enhanced aggregation of experts, is a one-step aggregation scheme that feeds directly the experts and the exogenous features to a unique nonlinear enhancer:

$$\hat{y} = h(\boldsymbol{\xi}, \boldsymbol{x}) \tag{21}$$

Of course, a variety of choices is possible for describing the nonlinear relationship $h(\cdot, \cdot)$. Here, an MLP-ANN is used and its structure, calibrated through 5-fold cross-validation, consists of one hidden layer with 20 neurons, hyperbolic tangent activation function and regularization term $\alpha = 0.01$.

For the benchmark problem, seven MLP-ANN's were trained, one for each weekday, fed by the vector $\boldsymbol{\xi} \in \mathbb{R}^{12}$ of experts predictions and the quarter-hour of the day t_{day}:

$$\hat{y} = h_{\mathrm{weekday}}(\boldsymbol{\xi}, t_{day}), \quad \mathrm{weekday} \in \{\mathrm{Monday}, \ldots, \mathrm{Sunday}\} \tag{22}$$

4. Results

In this Section the two proposed approaches, AEE and EAE, are applied to the German load benchmark. First the effects of the individual enhancers are

presented. Subsequently, the enhanced experts are aggregated according to the AEE scheme and the results compared to those of the EAE scheme and to the traditional state-of-the-art approaches, i.e. the aggregation of the plain experts.

4.1. Individual enhancement

For the benchmark problem, each individual enhancer can be visualized as seven surfaces, one for each day. The surfaces return the enhanced expert prediction, i.e. the enhanced load forecast, as a function of the original expert prediction and t_{day}. For a given weekday, displaying these surfaces offers an effective visualization of the alternative individual enhancement methods. In particular, let us consider the five enhancers discussed in Section 3.2. For expert 12, the five 'Thursday surfaces' are displayed in Fig. 6, together with the test data. The first plot is the surface corresponding to no enhancement, i.e. the 45-degree plane $\hat{\xi}_{12} = \xi_{12}$. In the insets, the corresponding Goodness of Fit plots are provided for the test data. The closer the surface is to the test data, the more effective is the enhancer. It can be seen that, without enhancement, $R^2 = 0.68$. All the five enhancers raise $R^2 =$ above 0.85, the maximum value being achieved by the affine and nonlinear tensor enhancers ($R^2 = 0.88$).

The MAPE and RMSE on data of each expert is compared to the ones of its five enhanced versions in Table 2 and Table 3. It appears that in all cases, individual enhancement significantly improves the base experts. Although no enhancer is uniformly superior to the others, the affine enhancer ranks first in seven cases out of twelve for both MAPE and RMSE. The second best enhancer appears to be the tensor spline one that ranks first in five cases out of twelve for both MAPE and RMSE.

The different flexibility of the alternative enhancers can be appreciated in Fig. 7, where three sections (at times 06:00, 06:30, and 07:00) of the five 'Tuesday surfaces' of the five enhancers are superimposed. In particular, it can be seen that the additive and multiplicative enhancers are not flexible enough to follow the training and data. Conversely, the other three enhancers offer a comparable performance.

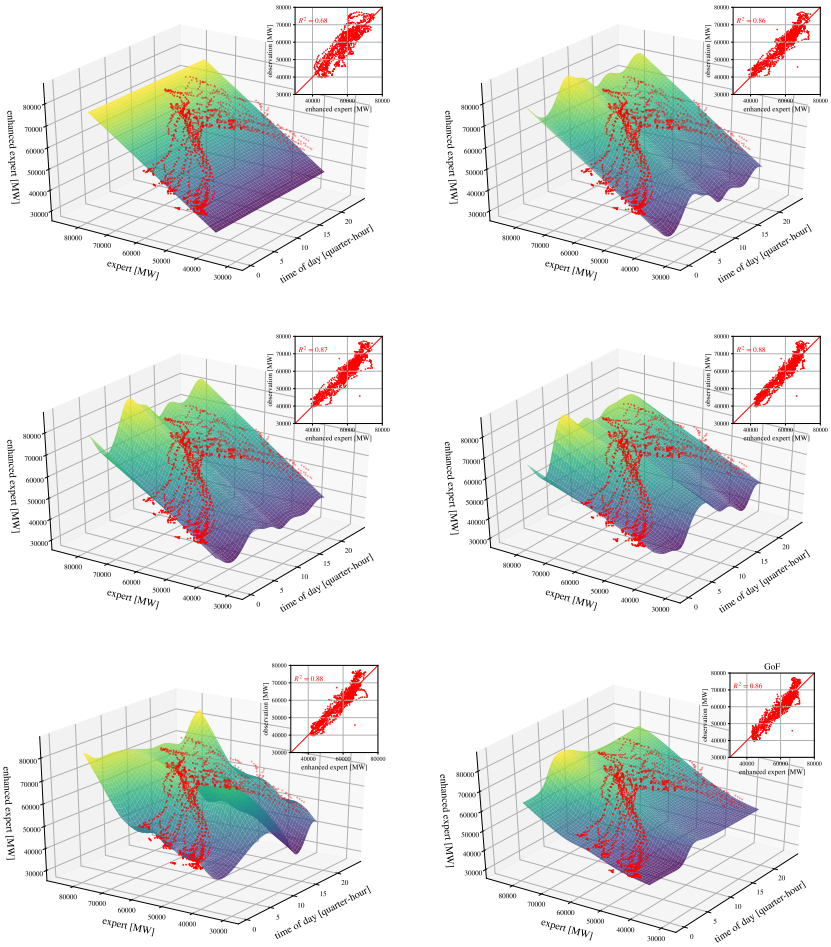

Figure 6: Expert 12, Thursday surfaces: actual loads against expert's forecast and time of day. Test data: red; enhancement function: surface. Insets: Goodness-of-Fit plots for test data. Top left panel: no enhancement, top right panel: additive enhancement, middle left panel: multiplicative enhancement, middle right panel: affine enhancement, bottom left panel: nonlinear tensor enhancement, bottom right panel: nonlinear MLP enhancement.

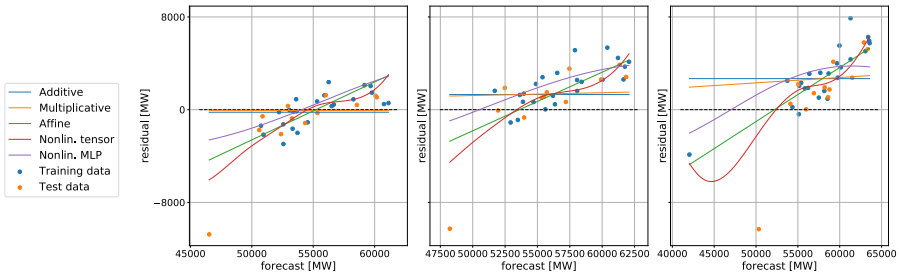

Figure 7: Expert 8: forecasting residuals against expert's forecast on Tuesday in correspondence of three times of day (from left panel to right panel: 06:00, 06:30, 07:00). Training data: blue, data: red, enhancement functions: continuous lines.

4.2. AEE and EAE results

For the AEE approach, 25 results were obtained, by considering all the possible pairs given by one of the five individual enhancers (Additive, Multiplicative, Affine, Nonlin. tensor, Nonlin. MLP-ANN) associated with one of the five aggregation methods (simple, trimmed, Winsorized, CLS, MLP). Conversely, the EAE approach yields a single result.

Two testing scenarios were considered: a 'full information' scenario, where at each time instant all the twelve experts are available, and a 'missing-experts' one, where on some days only a subset of the experts is available.

Full information scenario

For the German load benchmark, test MAPEs of the aggregated forecasts in the 'full information' scenario are summarized in Table 4. By a comparison with Table 2, the significant benefits brought by the two-stage strategies are apparent. While the original experts' MAPEs ranged from 2.80% to 8.61%, the AEE final MAPEs range from 1.95% to 2.86%, while the EAE scheme achieves a MAPE of 2.03%. Among the proposed methods, the Affine-CLS AEE method provides the best results, yielding a 30% improvement with respect to the best not-enhanced expert.

Table 2: Test MAPE [%] of each expert with and without enhancements, with the percentage decrease with respect to the not-enhanced expert below the best results.

Expert	**Enhancing methods**					
	None	**Add.**	**Mul.**	**Aff.**	**Tensor**	**MLP**
Expert 1	5.24	3.74	3.78	**3.38** (**-35%**)	3.48	3.68
Expert 2	4.92	3.76	3.82	3.51	**3.41** (**-31%**)	3.6
Expert 3	4.96	3.84	3.89	3.59	**3.52** (**-30%**)	3.75
Expert 4	4.91	3.82	3.87	3.55	**3.45** (**-30%**)	3.64
Expert 5	4.91	3.83	3.88	3.56	**3.43** (**-30%**)	3.65
Expert 6	4.17	3.28	3.33	**2.99** (**-28%**)	3.02	3.22
Expert 7	3.83	3.15	3.18	3.07	**3.03** (**-21%**)	3.21
Expert 8	3.48	2.63	2.66	**2.44** (**-30%**)	2.52	2.62
Expert 9	2.80	2.14	2.16	**2.01** (**-28%**)	2.23	2.15
Expert 10	5.33	2.83	2.84	**2.72** (**-49%**)	2.86	3.17
Expert 11	3.54	2.52	2.53	**2.42** (**-32%**)	2.84	2.67
Expert 12	8.61	5.31	5.16	**4.68** (**-46%**)	4.74	5.02

Concerning AEE, it can be noted that: (i) inspection of the best expert column shows that individual enhancement plays a key role in the final performance because the best enhanced experts is already close to the optimal performance; (ii) the choice of the aggregation method is crucial, with CLS and MLP outperforming the other three simpler alternatives. The test RMSEs, reported in Table 5, can be compared with Table 3. Considerations are similar to those already made for the MAPE results.

Table 3: Test RMSE [MW] of each expert with and without enhancements, with the percentage decrease with respect to the not-enhanced expert below the best results.

Expert	**Enhancing methods**					
	None	**Add.**	**Mul.**	**Aff.**	**Nonlin**	**MLP**
Expert 1	3634	2869	2883	**2683** (**-26%**)	2894	2805
Expert 2	3416	2862	2888	2718	**2662** (**-22%**)	2781
Expert 3	3436	2899	2924	2759	**2731** (**-21%**)	2846
Expert 4	3448	2930	2956	2772	**2745** (**-20%**)	2792
Expert 5	3435	2940	2966	2786	**2709** (**-21%**)	2831
Expert 6	3009	2548	2567	**2373** (**-21%**)	2475	2517
Expert 7	2854	2535	2547	2467	**2449** (**-14%**)	2567
Expert 8	2541	2098	2098	**1942** (**-24%**)	2115	2029
Expert 9	2102	1754	1749	**1658** (**-21%**)	2049	1744
Expert 10	3869	2291	2268	**2205** (**-43%**)	2353	2376
Expert 11	2737	2117	2108	**2029** (**-26%**)	3640	2033
Expert 12	5752	3766	3703	**3529** (**-39%**)	3584	3704

Fig. 8 displays one week of data, plotting the predictions provided by Affine-CLS AEE (red) and EAE (green) during the four days (Tuesday, Thursday, Friday, Sunday) randomly chosen for testing. The comparison with Fig. 4, where the predictions of two base experts were displayed, shows that both aggregation strategies AEE and EAE are capable to provide accurate predictions of the testing data.

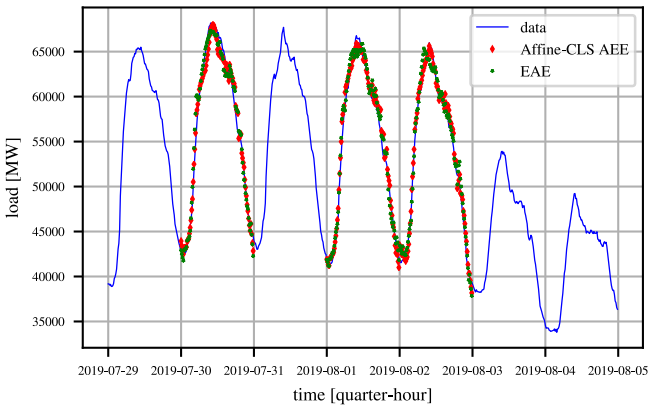

Figure 8: German load and predictions against time during one week.

Table 4: German data: test MAPE [%] of the proposed aggregation techniques. The best results are highlighted in bold; between parentheses, the percentage decrease with respect to the best not-enhanced expert (lower left corner).

		AEE					**EAE**
Individual Enhancement	Best expert	Simple average	Trimm. average	Winsor. average	Constr. Least Squares	**MLP**	
Additive	2.14	2.75	2.82	2.78	**2.09 (-25%)**	2.12	
Multiplicative	2.16	2.79	2.86	2.82	**2.11 (-25%)**	2.14	
Affine	2.01	2.53	2.57	2.54	**1.95 (-30%)**	2.03	**2.03 (-28%)**
Nonlin. tensor	2.23	2.50	2.51	2.48	**2.11 (-25%)**	2.23	
Nonlin. MLP-ANN	2.15	2.69	2.72	2.69	**2.00 (-29%)**	2.02	
No-enhancement	2.80	3.89	3.84	3.82	2.80	2.38	

Table 5: German data: test RMSE [MW] of the proposed aggregation techniques. The best results are highlighted in bold; between parentheses, the percentage decrease with respect to the best not-enhanced expert (lower left corner).

		AEE					**EAE**
Individual Enhancement	Best expert	Simple average	Trimm. average	Winsor. average	Constr. Least Squares	**MLP**	
Additive	1754	2179	2234	2203	**1714 (-18%)**	1745	
Multiplicative	1749	2190	2247	2213	**1708 (-19%)**	1761	
Affine	1658	2030	2072	2043	**1606 (-24%)**	1710	**1678 (-20%)**
Nonlin. tensor	2049	2102	2100	2076	**1911 (-9%)**	1966	
Nonlin. MLP-ANN	1744	2143	2201	2110	**1654 (-21%)**	1713	
No-enhancement	2102	2715	2702	2686	2081	1855	

Missing-experts scenario

In the missing-experts scenario, one, multiple, or even all experts may be unavailable during some time windows. It is important to consider this case as well, because these occurrences are not rare, due to disruptions in communications and software or hardware failures [23]. While mean-based aggregation techniques can naturally deal with missingness, weighted and nonlinear aggregation methods cannot. Herein, the conditional mean imputation method is adopted to deal with these phenomena without need to re-calibrate the aggregation models [24], [25]. The conditional mean imputation method uses the first and second order moments of the joint distribution of the experts, in order to derive linear imputation formulas predicting one or more experts from the knowledge of the others.

In order to test the robustness of the proposed aggregation strategies, in the test data 40% of the expert predictions were randomly labelled as missing. The resulting MAPEs are reported in Table 6. Again, the best results are obtained by Affine-CLS AEE. The AEE schemes achieve a MAPE ranging from 2.12% to 2.92%, while the EAE scheme achieves MAPE = 2.23%. As expected, MAPE and RMSE of both AEE and EAE increase with respect to the full information scenario. Nevertheless, both AEE and EAE, even when applied to the missing data scenario, are still able to improve over the best expert that uses all data, which demonstrates the remarkable robustness of the two enhanced aggregation strategies. The overall robustness of AEE and EAE is confirmed also when the RMSEs are considered, see Table 7.

Enhancement vs no-enhancement

A major question is whether the forecasting benefits are worth the effort of designing an enhanced aggregation scheme. In order to answer it, the last line of Tables 4-5 reports the MAPE and RMSE that are obtained if, without performing any enhancement, the experts are just aggregated according to standard methods. The improvement of AEE and EAE over No-Enhancement Aggregation (NEA) is very neat. For example in the full information case the

MAPE of the best NEA scheme (MLP) is 2.38% compared to 1.95% and 2.03% achieved by Affine-CLS AEE and EAE, respectively. Not only the best AEE and EAE schemes outperform NEA, but even most of non optimal AEE end EAE schemes provide a definite improvement. This superiority of enhanced schemes vs not-enhanced ones holds for both the full information and the missing-experts scenarios.

Table 6: German data: test MAPE [%] of the proposed aggregation techniques: missing-experts scenario. The best results are highlighted in bold; between parentheses, the percentage decrease with respect to the best not-enhanced expert that uses all data (lower left corner).

		AEE					**EAE**
Individual Enhancement	Best expert	Simple average	Trimm. average	Winsor. average	Constr. Least Squares	**MLP**	
Additive	2.30	2.83	2.88	2.84	**2.26 (-19%)**	2.34	
Multiplicative	2.31	2.88	2.92	2.89	**2.28 (-19%)**	2.36	
Affine	2.18	2.60	2.62	2.60	**2.12 (-24%)**	2.25	**2.23 (-20%)**
Nonlin. tensor	2.45	2.56	2.57	2.55	**2.27 (-19%)**	2.5	
Nonlin. MLP-ANN	2.45	2.72	2.76	2.77	**2.16 (-23%)**	2.21	
No-enhancement (full information)	2.80	3.91	3.87	3.85	2.95	2.63	

Table 7: German data: test RMSE [MW] of the proposed aggregation techniques: missing-experts scenario. The best results are highlighted in bold; between parentheses, the percentage decrease with respect to the best not-enhanced expert that uses all data (lower left corner).

		AEE					**EAE**
Individual Enhancement	Best expert	Simple average	Trimm. average	Winsor. average	Constr. Least Squares	**MLP**	
Additive	1876	2229	2273	2244	**1844 (-12%)**	1952	
Multiplicative	1873	2241	2288	2256	**1840 (-12%)**	1981	
Affine	1784	2072	2108	2080	**1732 (-18%)**	1879	**1911 (-9%)**
Nonlin. tensor	2248	2118	2123	2103	**1978 (-6%)**	2157	
Nonlin. MLP-ANN	1801	2144	2163	2146	**1772 (-16%)**	1811	
No-enhancement (full information)	2102	2728	2722	2702	2192	2101	

Causal scenario

Finally, Tables 8 and 9 report the results that are obtained when a (full information) strictly causal training-testing splitting is adopted, i.e. algorithms trained on data ranging from January 1 to July 31 and tested on the remaining load data. Since the training data do not cover an entire year, this scenario calls for some robustness against seasonal drifts. As it can be seen all the conclusions of the previous subsections are confirmed. In particular, the key role of individual enhancement, the superiority of CLS and MLP over other enhancers and the superiority of enhanced schemes vs not-enhanced ones.

Table 8: German data: test MAPE [%] of the proposed aggregation techniques: causal scenario. The best results are highlighted in bold; between parentheses, the percentage decrease with respect to the best not-enhanced expert (lower left corner).

		AEE					**EAE**
Individual Enhancement	Best expert	Simple average	Trimm. average	Winsor. average	Constr. Least Squares	**MLP**	
Additive	2.13	3.25	3.34	3.28	**2.07 (-32%)**	2.10	
Multiplicative	2.18	3.35	3.44	3.38	**2.10 (-31%)**	2.16	
Affine	2.06	2.87	2.89	2.85	**1.98 (-35%)**	2.07	**2.16 (-29%)**
Nonlin. tensor	2.26	2.81	2.84	2.79	**2.15 (-30%)**	2.27	
Nonlin. MLP-ANN	2.72	3.28	3.32	3.27	**2.52 (-18%)**	2.57	
No-enhancement	3.06	4.76	4.82	4.75	3.13	2.51	

Table 9: German data: test RMSE [MW] of the proposed aggregation techniques: causal scenario. The best results are highlighted in bold; between parentheses, the percentage decrease with respect to the best not-enhanced expert (lower left corner).

		AEE					**EAE**
Individual Enhancement	Best expert	Simple average	Trimm. average	Winsor. average	Constr. Least Squares	**MLP**	
Additive	1472	1972	2017	1987	**1402 (-27%)**	1431	
Multiplicative	1489	2007	2058	2024	**1409 (-27%)**	1422	
Affine	1437	1783	1800	1777	**1195 (-38%)**	1357	**1438 (-26%)**
Nonlin. tensor	1567	1767	1787	1760	**1451 (-25%)**	1501	
Nonlin. MLP-ANN	1764	2045	2070	2041	**1639 (-15%)**	1675	
No-enhancement	1932	2704	2745	2704	1946	1631	

5. Conclusions

In the time series forecasting literature, it is known that aggregating forecasts (herein called 'experts') coming from different external providers can significantly improve the final accuracy. However, possible drifts of the joint distribution of targets and features, associated with the difficulty of retuning the experts' models, can cause a drastical drop in the performances of the aggregated prediction.

In this work two general nonlinear approaches are proposed in order to overcome this issue. In both methods, the main novelty with respect to the existing literature is the introduction of some form of nonlinear 'enhancement', i.e. the exploitation of the exogenous features in order to fix errors and biases of the single experts. The first strategy is called Aggregation of Enhanced Experts (AEE) and is a two-stage method where the single experts are individually enhanced before being aggregated. The second strategy, Enhanced Aggregation of Experts (EAE) performs a unique nonlinear enhancement of the experts' predictions.

While in the present paper the effectiveness of the approach was demonstrated on the German load forecasting problem, the proposed methods can be easily generalized to a range of forecasting problems where the available experts need recalibration, but the underlying models cannot be retrained, possibly because the raw data are no more available or they are used as black-boxes, without having access to their inner structure. It is worth noting that, before proceeding with the enhancement process, a preliminary analysis of the expert performances is required in order to detect their shortcomings and select the best set of primary features to be used as inputs in the enhancement modules.

On the German load problem, AEE amd EAE achieved comparable results, achieving a $\sim 30\%$ reduction of the MAPE and $\sim 24\%$ reduction of the RMSE, compared to the performance of the best expert. In such a case, AEE could be preferable, because, especially if a linear aggregation is used, the prediction mechanism is much more transparent than EAE. In case of large prediction

errors, for instance, it is straightforward to isolate the failing experts, which could help the search for the root causes of the poor performance.

A further motivation for the development of smart aggregation techniques is the case of missing experts. Again with reference to the German load benchmark, a scenario where 40% of the experts are randomly missing was considered. It is remarkable that both AEE and EAE prove robust in this rather extreme case. In fact, AEE and EAE applied to the missing-experts scenario still guarantee a $19-24\%$ reduction of the MAPE and a $6-18\%$ reduction of the RMSE with respect to the best expert in the full information scenario.

The German load benchmark shows that the introduction of enhancement schemes can bring a definite advantage with respect the simple aggregation of the experts according to standard methods. For instance, in the full information case the MAPE passes from 2.38% for MLP aggregation without enhancement to 1.95% with AEE, corresponding to a neat 18% improvement.

In conclusion, smart aggregation strategies leveraging on some form of enhancement appear to be not only generally advantageous but also robust with respect to missing-experts scenarios.

Acknowledgements

This work has received funding from the European Research Council under the European Unions Horizon 2020 research and innovation programme (grant agreement no. 731232) and has been partially supported by the Italian Ministry for Research in the framework of the 2017 Program for Research Projects of National Interest (PRIN), Grant no. 2017YKXYXJ.

References

[1] M. Jaradat, M. Jarrah, A. Bousselham, Y. Jararweh, M. Al-Ayyoub, The internet of energy: Smart sensor networks and big data management for smart grid, Vol. 56, 2015. doi:10.1016/j.procs.2015.07.250.

[2] S. Yerpude, T. Singhal, Impact of internet of things (iot) data on demand forecasting, Indian Journal of Science and Technology 10. doi:10.17485/ijst/2017/v10i15/111794.

[3] J. Hox, H. Boeije, Data collection, primary versus secondary., Encyclopedia of Social Measurement 1. doi:10.1016/B0-12-369398-5/00041-4.

[4] G. I. Webb, Z. Zheng, Multistrategy ensemble learning: reducing error by combining ensemble learning techniques, IEEE Transactions on Knowledge and Data Engineering 16 (8) (2004) 980–991. doi:10.1109/TKDE.2004.29.

[5] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, Vol. 14, 2012. doi:10.1201/b12207.

[6] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140. doi:10.1007/BF00058655.

[7] D. Opitz, R. Maclin, Popular ensemble methods: An empirical study, Journal of Artificial Intelligence Research 11 (1999) 169–198. doi:10.1613/jair.614.

[8] J. Sill, G. Takacs, L. Mackey, D. Lin, Feature-weighted linear stacking (911). doi:10.1.1.379.8727.

[9] L. Wang, S. Mao, B. M. Wilamowski, R. M. Nelms, Ensemble learning for load forecasting, IEEE Transactions on Green Communications and Networking 4 (2) (2020) 616–628. doi:10.1109/TGCN.2020.2987304.

[10] J. Nowotarski, B. Liu, R. Weron, T. Hong, Improving short term load forecast accuracy via combining sister forecasts, Energy 98 (2016) 40–49. doi:10.1016/j.energy.2015.12.142.

[11] J. B. Predd, D. N. Osherson, S. R. Kulkarni, H. V. Poor, Aggregating Probabilistic Forecasts from Incoherent and Abstaining Experts, Decision Analysis 5 (4) (2008) 177–189. doi:10.1287/deca.1080.0119.

[12] B. Turner, M. Steyvers, E. Merkle, D. Budescu, T. Wallsten, Forecast aggregation via recalibration, Machine Learning doi:10.1007/s10994-013-5401-4.

[13] F. Bolger, G. Rowe, The aggregation of expert judgment: Do good things come to those who weight?, Risk Analysis 35 (1) (2015) 5–11. doi:https://doi.org/10.1111/risa.12272.

[14] P. Gaillard, Y. Goude, R. Nedellec, Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting, International Journal of Forecasting 32 (3) (2016) 1038–1050. doi:10.1016/j.ijforecast.2015.

[15] K. F. Wallis, Combining forecasts – forty years later, Applied Financial Economics 21 (1-2) (2011) 33–41. doi:10.1080/09603107.2011.523179.

[16] R. Ranjan, T. Gneiting, Combining probability forecasts, Journal of the Royal Statistical Society Series B 72 (2010) 71–91. doi:10.2307/40541575.

[17] M. Devaine, P. Gaillard, Y. Goude, G. Stoltz, Forecasting electricity consumption by aggregating specialized experts doi:10.1007/s10994-012-5314-7.

[18] Entso-e transparency platform, https://transparency.entsoe.eu/.

[19] T. Hastie, R. Tibshirani, Generalized additive models: Some applications, Journal of the American Statistical Association 82 (398) (1987) 371–386. doi:10.1007/978-1-4615-7070-7_8.

[20] A. Pierrot, Y. Goude, Short-term electricity load forecasting with generalized additive models, 2011.

[21] D. Kingma, J. Ba, Adam: A method for stochastic optimization, International Conference on Learning Representations.

[22] V. Genre, G. Kenny, A. Meyler, A. Timmermann, Combining expert forecasts: Can anything beat the simple average?, International Journal of

Forecasting 29 (2013) 108–121. doi:10.1016/j.ijforecast.2012.06.004.

[23] C. Fraley, A. Raftery, T. Gneiting, Calibrating multimodel forecast ensembles with exchangeable and missing members using bayesian model averaging, Monthly Weather Review - MON WEATHER REV 138. doi:10.1175/2009MWR3046.1.

[24] E. M. L. Beale, R. J. A. Little, Missing values in multivariate analysis, Journal of the Royal Statistical Society: Series B (Methodological) 37 (1) (1975) 129–145. doi:10.1111/j.2517-6161.1975.tb01037.x.

[25] S. Buck, A method of estimation of missing values in multivariate data suitable for use with an electronic computer, Journal of the Royal Statistical Society. Series B 22. doi:10.1111/j.2517-6161.1960.tb00375.x.

[26] G. Wahba, Spline Models for Observational Data, Society for Industrial and Applied Mathematics, Philadelphia, 1990. doi:10.1137/1.9781611970128.

[27] H. Prautzsch, W. Boehm, M. Paluszny, Bezier and B-Spline Techniques, Springer-Verlag, Berlin, Heidelberg, 2002.

[28] P. H. C. Eilers, B. D. Marx, Flexible smoothing with b-splines and penalties, STATISTICAL SCIENCE 11 (1996) 89–121. doi:10.1214/SS/1038425655.

[29] C. d. Boor, A Practical Guide to Splines, Springer Verlag, New York, 1978. doi:10.2307/2006241.

Appendix

Cyclic cubic P-splines

Splines are special functions able to model nonlinear behaviour while avoiding overfitting and boundary issues which are typical of high-order polynomials

in regression problems [26]. They are obtained by choosing a set of control points, called 'knots', that split the dataset into bins, and fitting within each interval a low-order polynomial, while applying continuity constraints at the knots. Cubic splines in particular are piecewise cubic polynomials with continuous derivatives up to order 2 at each knot.

Let consider the observations y_i, and the corresponding predictors x_i, $i = 1, ..., N$. The regression cubic spline $\hat{f}$ is the solution of the following optimization problem:

$$\hat{f} = \arg\min_{\tilde{f}} \sum_{i=1}^{N} \left(y_i - \tilde{f}(x_i) \right)^2, \quad \tilde{f} = \sum_{j=1}^{K} \beta_j h_j(x) \tag{23}$$

where $\tilde{f}$ is a cubic spline, β_j are the spline coefficients, K is the number of knots and $h_j(x)$ are the basis functions. Although there is no unique choice for the basis functions, the B-splines are widely used in the literature because of their computational properties [27].

There is a variety of strategies for enforcing smoothness to the fitted curve, for example reducing the number of knots. A typical choice is to keep a high number of knots and add a penalty term to the previous optimization problem, which becomes

$$\hat{f} = \arg\min_{\tilde{f}} \sum_{i=1}^{N} \left(y_i - \tilde{f}(x_i) \right)^2 + \lambda \int_{-\infty}^{\infty} \left(\tilde{f}''(x) \right)^2 dx \tag{24}$$

where λ is the regularization parameter.

The result of the above optimization is the cubic penalized B-spline (cubic P-spline) [28]. An additional constraint can be applied to the spline basis in order to enforce continuity between the first and the last boundary of the considered independent variable. This leads to the cyclic cubic P-splines, which is useful whenever it is necessary to model periodic behaviours within the considered variables (e.g. day of year or time of day) [29]. In the present paper, cyclic cubic P-splines are used to model the effect of the time of the day. Given that data are sampled on a quarter-hourly period, the effect is modelled by a periodic function of period 96 on the quarter-hour time scale.

Authors

Alessandro Incremona is currently a Post-Doctoral researcher in Electronics, Computer Science and Electrical Engineering at the University of Pavia. He was a student of the Almo Collegio Borromeo of Pavia, and received the Bachelor and Master degrees (with honor) in Computer Engineering (Embedded and Control Systems) in 2015 and 2017, respectively. In 2021 he received the PhD degree in Electronics, Electrical and Computer Engineering from the University of Pavia. From May to August 2019 he was a Research Intern with IBM Research, Dublin, Ireland. His research interest are: time series forecasting, machine learning, statistical process control, energy forecasting, optimization and regularization.

Giuseppe De Nicolao has a long term experience in system identification, Model Predictive Control, statistical methods, neural networks, and their application to modelling and control of industrial and biomedical systems. Full Professor of Model Identification and Data Analysis at the University of Pavia. Keynote speaker at the 1st Workshop on Assessment and Future Direction of NMPC (1998). Author/coauthor of more than 140 papers in peer reviewed journals. Associate Editor of Automatica and former Associate Editor of the IEEE Trans. on Automatic Control and IEEE Transactions on Control Systems Technology.

Francesco Fusco received the master degree in industrial automation engineering from Università Politecnica delle Marche (UNIVPM), Ancona, Italy, in 2008. He completed his PhD Degree in electronic engineering at the National University of Ireland Maynooth (NUIM), Ireland,

in 2012, with his work on forecasting and control of wave energy converters. Since 2012, he has been a Research Staff Member with IBM Research Europe, Ireland, where he matured expertise in the development of large-scale data analytics and machine learning systems, with particular focus on electricity forecasting.

Bradley J. Eck is research scientist and manager at IBM Research Europe in Dublin, Ireland. His research interests include machine learning for industrial applications, with a special focus on physical processes and systems. Output of this work comprise over 40 peer reviewed publications and several open source software packages. Brad's experience also includes managing commercial and state-funded initiatives to demonstrate new technology in practical settings. Dr Eck holds a PhD in Civil Engineering from the University of Texas at Austin.

Seshu Tirupathi is a Research Scientist in IBM Research Dublin since 2014. His current research interests include applications of machine learning algorithms, cloud computing, numerical methods and scientific computing. Over the years in IBM, Seshu has made significant contributions to multiple domains like ground water, surface water, electricity networks through fundamental research as well as consulting projects. Prior to joining IBM, Seshu earned his PhD in Applied Mathematics from Brown University. He also holds an M.Eng. in Mechanical Engineering from Cornell University and a B.Tech. degree in Aerospace Engineering from Indian Institute of Technology Kanpur.