ISSN 1440-771X

MONASH University

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Probabilistic time series forecasting with boosted additive models: an application to smart meter data

Souhaib Ben Taieb, Raphael Huser,
Rob J Hyndman, Marc G Genton

June 2015

Working Paper 12/15

Probabilistic time series forecasting with boosted additive models: an application to smart meter data

Souhaib Ben Taieb
CEMSE division
King Abdullah University of Science and Technology
Thuwal 23955-6900, Saudi Arabia
souhaib.bentaieb@kaust.edu.sa

Raphael Huser
CEMSE division
King Abdullah University of Science and Technology
Thuwal 23955-6900, Saudi Arabia
raphael.huser@kaust.edu.sa

Rob J. Hyndman
Department of Econometrics and Business Statistics
Monash University
VIC 3800, Australia
rob.hyndman@monash.edu

Marc G. Genton
CEMSE division
King Abdullah University of Science and Technology
Thuwal 23955-6900, Saudi Arabia
marc.genton@kaust.edu.sa

9 June 2015

JEL classification: Q47, C14, C22

Probabilistic time series forecasting with boosted additive models: an application to smart meter data

Abstract

A large body of the forecasting literature so far has been focused on forecasting the conditional mean of future observations. However, there is an increasing need for generating the entire conditional distribution of future observations in order to effectively quantify the uncertainty in time series data. We present two different methods for probabilistic time series forecasting that allow the inclusion of a possibly large set of exogenous variables. One method is based on forecasting both the conditional mean and variance of the future distribution using a traditional regression approach. The other directly computes multiple quantiles of the future distribution using quantile regression. We propose an implementation for the two methods based on boosted additive models, which enjoy many useful properties including accuracy, flexibility, interpretability and automatic variable selection. We conduct extensive experiments using electricity smart meter data, on both aggregated and disaggregated scales, to compare the two forecasting methods for the challenging problem of forecasting the distribution of future electricity consumption. The empirical results demonstrate that the mean and variance forecasting provides better forecasts for aggregated demand, while the flexibility of the quantile regression approach is more suitable for disaggregated demand. These results are particularly useful since more energy data will become available at the disaggregated level in the future.

Keywords: Additive models; Boosting; Density forecasting; Energy forecasting; Probabilistic forecasting.

1 Introduction

Optimal decision making in critical problems of science and society requires the quantification of uncertainty for future events (Spiegelhalter, 2014). In particular, probabilistic forecasting of a time series involves generating probability distributions for future observations at multiple forecast horizons using a sequence of historical observations, possibly considering additional

related time series. Probabilistic forecasting (Gneiting and Katzfuss, 2014) has become increasingly important in various fields, including healthcare (Jones and Spiegelhalter, 2012), climate science (Palmer, 2012) and finance (Groen, Paap, and Ravazzolo, 2013). However, a large part of the forecasting literature has focused on point forecasting, i.e., forecasting only the mean or median of the future distribution. As a result, there is a need for the design and evaluation of different probabilistic forecasting methods that can provide not only point forecasts but, more importantly, quantify the uncertainty in future time series data.

Recently, the energy sector has been changing dramatically, notably due to the integration of renewable energy sources, as an effort to reduce our dependency on fossil fuels and achieve a better sustainable future. With the growing amount of data from energy systems, there is a need from the utilities to generate probabilistic forecasts, especially for wind power (Zhu and Genton, 2012), solar power (Bacher, Madsen, and Nielsen, 2009) and electricity demand (Wijaya, Sinn, and Chen, 2015). In particular, accurate probabilistic forecasts for electricity demand is critical for electric utilities in many operational and planning tasks.

Electricity load is often represented as the aggregated load across many households (e.g., at the city level). There is also a rich literature on forecasting the average aggregated electricity load (Alfares and Nazeeruddin, 2002), i.e., in forecasting the mean of the future demand distribution. These forecasts are often generated conditional on a number of predictor variables such as calendar and temperature variables. Various models have been considered for modeling and forecasting the average electricity load, including linear regression (Hong, 2010), (seasonal) ARMA models (Taylor, 2010), neural networks (Hippert, Pedreira, and Souza, 2001) and (boosted) additive models (Ben Taieb and Hyndman, 2014; Fan and Hyndman, 2012). However, the literature on probabilistic load forecasting is relatively sparse (see Tao and Fan, 2014, for a recent review).

In this article, we focus on the problem of probabilistic forecasting for smart meter data. A *smart meter* (Zheng, Gao, and Lin, 2013) is an electronic device that records and transmits electricity consumption information at 30-minute or hourly intervals, hence generating a huge quantity of data. Compared to traditional electricity load, smart meters measure the load at a very local level, typically for individual households. Because smart meter data are highly volatile, forecasting the average load does not provide meaningful information about the uncertainty of the future demand. Instead, we need to forecast the entire distribution of the future demand. In

other words, a shift is occurring from point forecasting to probabilistic forecasting (Gneiting and Katzfuss, 2014).

However, the literature on probabilistic forecasting for smart meter data is even more limited than for traditional electricity load forecasting. The only article we are aware of is Arora and Taylor (2014) who considered kernel density estimation methods to generate probabilistic forecasts for individual smart meters. One of the contributions we make in this paper is to enrich the literature on probabilistic forecasting for smart meter data.

Among the different models that have been considered in the literature for electricity demand forecasting, additive models (Fan and Hyndman, 2012) have been increasingly popular due to their accuracy, flexibility and interpretability. Boosted additive models are even more attractive since they provide a more flexible modeling procedure including automatic variable selection, as well as a better resistance to overfitting.

In this article, we make the following contributions:

1. We present two different methods for probabilistic time series forecasting which allow the inclusion of exogenous variables (Section 2). One method is based on forecasting both the conditional mean and variance of the future distribution using a traditional regression approach. The other directly computes multiple quantiles of the future distribution using quantile regression.

2. We propose an implementation of the two methods based on boosted additive models. In particular, we show how to fit the different (quantile) regression models involved for the two approaches (Section 3).

3. We consider the challenging problem of modeling and forecasting electricity smart meter data. In particular, our experiments are based on 250 meters (with more than 25000 observations for each meter) from a recently released public smart meter data set (Section 4).

4. We compare the two forecasting methods (including two additional benchmark methods) on all the 250 meters using different forecast accuracy measures. The different methods are compared on both disaggregated and aggregated electricity demand obtained by summing the demand of all the meters (Section 5).

2 Probabilistic forecasting

2.1 Problem formulation

In a wide range of applications, ranging from the energy sector to finance or climate science, it is of interest to estimate the future probability distribution of a given variable from historical data, possibly using a set of exogeneous variables. Mathematically, the problem may be formulated as follows: given a time series $\{y_1, \dots, y_T\}$ comprising T observations, the goal is to estimate the distribution $F_{t,h}$ of y_{t+h}, for forecast horizon $h = 1, \dots, H$, given the information available up to time t.

The problem of probabilistic forecasting can be reduced to the estimation of the conditional distribution

$$F_{t,h}(y \mid \boldsymbol{x}_t) = P(y_{t+h} \leq y \mid \boldsymbol{x}_t),$$

for any $h = 1, \dots, H$ and forecast time origin t, where

- $\boldsymbol{x}_t = (\boldsymbol{y}_t, \boldsymbol{z}_{t+h})$;
- $\boldsymbol{y}_t$ is a vector of lagged observations prior to and including time t;
- $\boldsymbol{z}_{t+h}$ is a vector of exogeneous variables prior to and including time $t+h$.

To this end, a popular approach is to forecast the conditional first and second moments and then make distributional assumptions to recover the entire distribution; see Section 2.2. Another more flexible possibility is to directly forecast conditional quantiles, and then combine them to produce a valid distribution function; see Section 2.3.

2.2 Conditional mean and variance forecasting

We may model the observations from time t to $t+h$ using a different model at each forecast horizon h as follows:

$$y_{t+h} = g_h(\boldsymbol{x}_t) + \sigma_h(\boldsymbol{x}_t)\varepsilon_{t+h}, \tag{1}$$

where $g_h(\cdot)$ and $\sigma_h(\cdot)$ denote smooth functions of predictors $\boldsymbol{x}_t = (\boldsymbol{y}_t, \boldsymbol{z}_{t+h})$, and $\{\varepsilon_t\}$ is an independent and identically distributed (i.i.d.) noise process with $\mathbb{E}[\varepsilon_t] = 0$ and $\mathbb{E}[\varepsilon_t^2] = 1$. Under this model, the conditional mean and variance of y_{t+h} are given by:

$$\mathbb{E}[y_{t+h} \mid \boldsymbol{x}_t] = g_h(\boldsymbol{x}_t), \tag{2}$$

$$\mathbb{V}[y_{t+h} \mid \boldsymbol{x}_t] = \sigma_h(\boldsymbol{x}_t). \tag{3}$$

In other words, the available information at time t, namely $\boldsymbol{x}_t$, characterizes the expectation and variance of y_{t+h}, provided $g_h(\cdot)$ and $\sigma_h(\cdot)$ are not constant.

Since we have $\mathbb{E}[(y_{t+h} - g_h(\boldsymbol{x}_t))^2 \mid \boldsymbol{x}_t] = \sigma_h^2(\boldsymbol{x}_t)$, one simple way to estimate the conditional mean $g_h(\boldsymbol{x}_t)$ and variance $\sigma_h^2(\cdot)$ proceeds as follows:

1. Apply a mean regression using $\{(y_{t+h}, \boldsymbol{x}_t)\}$ to obtain $\hat{g}_h(\boldsymbol{x}_t)$, an estimate of the conditional mean;
2. Compute empirical residuals $e_{t+h} = y_{t+h} - \hat{g}_h(\boldsymbol{x}_t)$;
3. Apply a mean regression using $\{(e_{t+h}^2, \boldsymbol{x}_t)\}$ to obtain $\hat{\sigma}_h^2(\cdot)$, an estimate of the conditional variance. To avoid negative values for e_{t+h}^2, we can use $(\log(e_{t+h}^2), \boldsymbol{x}_t)$ and backtransform using the exponential (see Chen, Cheng, and Peng, 2009).

For the model given in (1), we have made assumptions only about the first two moments of y_{t+h}. Without making further assumptions, we can use these two quantities to make probabilistic statements regarding deviations of observations around their mean value, for example using Chebyshev's inequality. One the other hand, if we make a normality assumption for y_{t+h}, then we can write

$$y_{t+h} \mid \boldsymbol{x}_t \sim \mathcal{N}(\hat{g}_h(\boldsymbol{x}_t), \hat{\sigma}_h^2(\boldsymbol{x}_t)), \tag{4}$$

where $\mathcal{N}(a, b)$ is a normal distribution with mean a and variance b, and a 95% prediction interval for y_{t+h} may be given by $\hat{g}_h(\boldsymbol{x}_t) \pm 1.96\hat{\sigma}_h(\boldsymbol{x}_t)$.

Thanks to the Central Limit Theorem, the normality assumption is often verified when variables can be seen as sums of many smaller components (e.g., with aggregated data). If the observations are indeed normally distributed, then this approach will likely provide the best forecasts. However, since we are now making assumptions about all the moments of the distribution, not just the first two moments, we might obtain poor forecasts of the tail quantiles if the observations are not normally distributed.

The approach described above has been considered in a linear regression context in Engle et al. (1993), with non-parametric regression in Fan and Yao (1998), and with additive models in Wijaya, Sinn, and Chen (2015).

2.3 Quantile forecasting

Instead of making assumptions about the form of the conditional distribution, we can use a more general approach where we compute the conditional τ-quantiles of the distribution for a set of Q probabilities τ_i, $i = 1, \ldots, Q$, e.g., $\tau_i = i/100$ with $Q = 99$. This can be achieved by moving from mean regression to quantile regression (Kneib, 2013; Koenker, 2005). Then, we can recover the predictive distribution from these quantiles (e.g., using linear interpolation after suitable adjustments to avoid quantile crossings), provided a large set of quantiles are computed.

The quantile regression model for the τ-quantile at forecast horizon h may be written as

$$y_{t+h} = g_{h,\tau}(\boldsymbol{x}_t) + \varepsilon_{t+h,\tau}, \quad F_{\varepsilon_{t+h},\tau}(0) = \tau, \tag{5}$$

where $F_{\varepsilon_{t+h},\tau}$ denotes the cumulative distribution function of $\varepsilon_{t+h,\tau}$, and where the smooth functions $g_{h,\tau}(\cdot)$ are distinct for each quantile and horizon.

Compared to the model in (1), the assumption of zero means for the error terms is replaced by the assumption of zero τ-quantiles. This implies that the conditional τ-quantile, issued at time t for lead time $t+h$ can be computed as follows:

$$q_{t,h}^{(\tau)} = F_{t,h}^{-1}(\tau \mid \boldsymbol{x}_t) = g_{h,\tau}(\boldsymbol{x}_t).$$

It is well-known that the conditional expectation may be estimated by minimizing the expected square loss, and that the conditional median may be estimated by minimizing the expected absolute loss. Similarly, we can show that the conditional τ-quantile $q_{t,h}^{(\tau)}$ can be computed using the pinball loss function (Gneiting, 2011), i.e.,

$$q_{t,h}^{(\tau)} = \arg\min_{q} \mathbb{E}[L_\tau(Y, q) \mid \boldsymbol{x}_t], \tag{6}$$

where the expectation is taken with respect to $Y \sim F_{t,h}$ and the pinball loss is defined as

$$L_\tau(y, q) = \begin{cases} \tau(y - q) & \text{if } y \geq q; \\ -(1-\tau)(y - q) & \text{if } y < q. \end{cases} \tag{7}$$

Notice that when $\tau = 0.5$, the pinball loss is equivalent to the absolute loss since $2L_\tau(y, q) = |y - q|$. Furthermore, the empirical counterpart $\hat{q}_{t,h}^{(\tau)}$ of (6) may be used for consistent estimation of $q_{t,h}^{(\tau)}$.

In order to produce a valid cumulative distribution function at horizon h, quantile forecasts need to satisfy the following monotonicity property:

$$\forall \tau_1, \tau_2 \text{ such that } \tau_1 \leq \tau_2 : \hat{q}_{t,h}^{(\tau_1)} \leq \hat{q}_{t,h}^{(\tau_2)}.$$

However, since each τ-quantile is modeled and estimated independently for each probability τ, the monotonicity property might be not satisfied for all quantiles; the problem is known as *quantile crossing* (Koenker, 2005). The number of quantile crossings will typically depend on the model complexity, the sample size, the number of estimated quantiles, and the distance between the chosen probabilities τ_i.

Multiple approaches have been proposed to deal with the problem of quantile crossing, including joint estimation or monotone rearranging (Chernozhukov, Fernández-Val, and Galichon, 2010); the latter is the approach that we adopt in this work.

2.4 Probabilistic forecast evaluation

Given a predicted cumulative distribution function $\hat{F}_{t,h}$, and an actual observation y_{t+h} at horizon h, we can evaluate the forecasting error using the continuous ranked probability score (CRPS) (Gneiting and Raftery, 2007), which can be defined equivalently as follows:

$$\text{CRPS}(\hat{F}_{t,h}, y_{t+h}) = \int_{-\infty}^{\infty} \left(\hat{F}_{t,h}(z) - \mathbb{1}(z \geq y_{t+h})\right)^2 dz \tag{8}$$

$$= 2 \int_0^1 L_\tau\left(y_{t+h}, \hat{F}_{t,h}^{-1}(\tau)\right) d\tau \tag{9}$$

$$= \underbrace{\mathbb{E}_{\hat{F}}|Y - y_{t+h}|}_{\text{Reliability}} - \underbrace{\frac{1}{2}\mathbb{E}_{\hat{F}}|Y - Y'|}_{\text{Spread}}, \tag{10}$$

where $\hat{F}$ is a shorthand for $\hat{F}_{t,h}$, and Y and Y' are two independent random variables with distribution $\hat{F}_{t,h}$.

Compared to other forecast accuracy measures, such as the probability integral transform (PIT), the CRPS quantifies both the *reliability* and *spread* of the probabilistic forecast; see (10). Reliability, sometimes also called *calibration*, measures the correspondence between the forecasts and the observations. In other words, a forecast is well-calibrated if there is a good match between forecasted and observed medians. The spread measures the lack of *sharpness* or *concentration* of the predictive distributions and is a property of the forecasts only. The best

forecast is the one that minimizes the spread (i.e., maximizes the sharpness) of the predictive distributions subject to calibration (Gneiting, Balabdaoui, and Raftery, 2007; Gneiting and Raftery, 2007).

Given a testing set of size N, we can compute the average CRPS for an h-step-ahead forecast as follows:

$$\text{CRPS}(h) = \frac{1}{N}\sum_{t=1}^{N}\text{CRPS}(\hat{F}_{t,h}, y_{t+h}). \tag{11}$$

In this work, we use the definition given in expression (10) since it allows us to decompose the CRPS into the reliability and spread components. More precisely, we approximate the CRPS at horizon h as follows:

$$\text{CRPS}(h) \approx \frac{1}{NM}\sum_{t=1}^{N}\left(\sum_{i=1}^{M}|y_{t+h}^{(i)} - y_{t+h}| - \frac{1}{2}\sum_{i=1}^{M}|y_{t+h}^{(i)} - y_{t+h}^{(i)'}|\right) \tag{12}$$

$$= \underbrace{\left(\frac{1}{NM}\sum_{t=1}^{N}\sum_{i=1}^{M}|y_{t+h}^{(i)} - y_{t+h}|\right)}_{\text{Estimated reliability}} - \underbrace{\left(\frac{1}{2NM}\sum_{t=1}^{N}\sum_{i=1}^{M}|y_{t+h}^{(i)} - y_{t+h}^{(i)'}|\right)}_{\text{Estimated spread}}, \tag{13}$$

where $y_{t+h}^{(i)}$ and $y_{t+h}^{(i)'}$ are two independent samples from the predictive distribution $\hat{F}_{t,h}$, and M is the number of random samples from $\hat{F}_{t,h}$.

We can also evaluate an h-step-ahead quantile forecast $\hat{q}_{t,h}^{(\tau)}$ with nominal proportion τ, by averaging the pinball losses over the whole testing set for the quantile τ. In other words, we can compute the average quantile loss as follows:

$$\text{QL}(h,\tau) = \frac{1}{N}\sum_{t=1}^{N}L_\tau(y_{t+h}, \hat{q}_{t,h}^{(\tau)}),$$

where L_τ is the pinball loss defined in (7).

Finally, a quantile forecast $\hat{q}_{t,h}^{(\tau)}$ for horizon h with nominal proportion τ can also be evaluated using the *unconditional coverage* $\tilde{\tau}^{(h)}$, which measures the percentage of observations that are lower than the forecasted τ-quantile at horizon h, i.e.,

$$\tilde{\tau}^{(h)} = \frac{1}{N}\sum_{t=1}^{N}\mathbb{1}(y_{t+h} \leq \hat{q}_{t,h}^{(\tau)}), \tag{14}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Ideally, a quantile forecast with a nominal level τ should satisfy $\tilde{\tau}^{(h)} = \tau$ for each horizon $h = 1, \ldots, H$. We can also evaluate the average unconditional coverage over the forecast horizons, i.e.

$$\tilde{\tau} = \frac{1}{H} \sum_{h=1}^{H} \tilde{\tau}^{(h)}, \tag{15}$$

which should also ideally satisfy $\tilde{\tau} = \tau$.

3 Boosted additive models

In practice, the functions $g_h(\cdot)$ and $\sigma_h(\cdot)$ in (1), and $g_{h,\tau}(\cdot)$ in (5) need to be estimated from data. For both mean and quantile regression, we can use a wide range of class of models, for example, linear models and neural networks (Hastie, Tibshirani, and Friedman, 2008). In this work, we will consider *boosted additive models* since they allow for flexibility, interpretability and automatic variable selection. Furthermore, unlike standard linear regression, boosted additive models can handle a large number of input variables of different types, while avoiding gross overfitting, which is particularly important for prediction.

3.1 Generalized Additive Models

Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1990) may be expressed as

$$y_{t+h} = \ell(m(\boldsymbol{x}_t)) = \beta_0 + a_1(x_{1t}) + \cdots + a_p(x_{pt}), \tag{16}$$

where ℓ is a link function (e.g., the identity or logarithm function), m is the function g_h, σ_h or $g_{h,\tau}$, the terms $a_k(\cdot)$ denote functions that may be specified parametrically or smooth functions estimated non-parametrically (e.g., using cubic splines), and x_{kt} is the kth component of the vector $\boldsymbol{x}_t$ with $k = 1, \ldots, p$.

One attractive feature of GAMs is their flexibility: since each function a_k can possibly have a non-linear shape, GAMs benefit from a high flexibility and can provide higher accuracy than, e.g., simple linear models. Another appealing feature is their interpretability: compared to full complexity models such as neural networks, GAMs can be easily interpreted as each function $a_k(x_{kt})$ is a function of a single variable and may be plotted against its input x_{kt}.

However, because standard GAMs do not model any interactions between the input variables, they can suffer from a low performance compared to full complexity models when such interactions truly exist. In any case, standard GAMs have been successfully applied in many studies for electricity demand forecasting (Cho et al., 2013; Fan and Hyndman, 2012). Extensions have been proposed to allow interactions in GAMs (Lou et al., 2013).

Backfitting (Hastie and Tibshirani, 1990) and gradient boosting (Friedman, 2001) are the two popular methods for fitting GAMs. Bühlmann and Yu (2003) has shown that the boosting procedure is competitive with respect to backfitting and can even outperform the latter in high dimensions p. In this work, we estimate GAMs using a gradient boosting procedure; in other words, we learn *boosted additive models* rather than *backfitted additive models*.

3.2 The gradient boosting algorithm

Boosting is a learning algorithm stemming from the machine learning literature based on the idea of creating an accurate learner by combining many so-called "weak learners" (Schapire, 1990), i.e., with high bias and small variance. Since its inception (Schapire, 1990), boosting has attracted much attention due to its excellent prediction performance in a wide range of applications both in the machine learning and statistics literatures (Schapire and Freund, 2012). Gradient boosting is a popular approach which interprets boosting as a method for function estimation from the perspective of numerical optimization in a function space (Friedman and Hastie, 2000; Friedman, 2001).

Given a dataset $\mathcal{D} = \{(y_{t+h}, \boldsymbol{x}_t)\}_{t=1}^{T}$ where y_t and $\boldsymbol{x}_t$ are linked through (16), and a loss function $L(y, m)$, the goal is to fit the model (16) by minimizing the loss L over the dataset $\mathcal{D}$. In the following, we will present the gradient boosting procedure for mean regression (also called L_2Boost (Bühlmann and Yu, 2003)) and quantile regression (also called quantile boosting (Mayr, Hothorn, and Fenske, 2012)). These correspond respectively to using $L_2(y, m) = (y - m)^2$ and L_τ defined in (7) as loss functions.

Denote by $\hat{\boldsymbol{m}}^{[j]} = \left(\hat{m}^{[j]}(\boldsymbol{x}_t)\right)_{t=1,\dots,T}$ the vector of function estimates at iteration $j = 1, \dots, J$, where J is the number of boosting iterations. The different steps of the gradient boosting algorithm can be written as follows:

1. Initialize the function estimate $\hat{\boldsymbol{m}}^{[0]}$ with starting values. The unconditional mean and the τth sample quantile are natural choices for mean regression and quantile regression,

respectively. The median has also been suggested as a starting value for quantile regression (Mayr, Hothorn, and Fenske, 2012).

2. Specify a set of B base-learners, and set $j = 0$. Base-learners are simple regression estimators (or weak learners) that depend on a subset of the initial set of input variables $\boldsymbol{x}_t$, and a univariate response. However, since we are fitting an additive model, each base-learner will depend on exactly one input variable.

   In this work, we will consider one base-learner with a linear effect for each categorical variable, and two base-learners for each continuous variable, with both a linear and a nonlinear effect. By doing so, we allow the boosting procedure to decide automatically if the nonlinear extension is required or if the linear effect is sufficient. In other words, given p input variables with c categorical variables, we will have a total of $B = 2 \times p - c$ base-learners.

   Note that for each boosting iteration, one of the base-learners will be selected. So, the final model will typically include only a subset of the initial variables.

3. Increase the number of iterations j by 1.

4. (a) Compute the negative gradient of the loss function evaluated at the function estimate of the previous iteration $\hat{\boldsymbol{m}}^{[j-1]}$:

$$\boldsymbol{u}^{[j]} = \left(-\frac{\partial}{\partial m} L(y_{t+h}, m)\Big|_{m=\hat{m}^{[j-1]}(\boldsymbol{x}_t)} \right)_{t=1,\ldots,T}$$

   For mean regression, that is with the L_2 loss function, the negative gradients are given by:

$$\boldsymbol{u}^{[j]} = \left(-2\left(y_{t+h} - \hat{m}^{[j-1]}(\boldsymbol{x}_t) \right) \right)_{t=1,\ldots,T}$$

   For quantile regression, that is with the L_τ loss function, the negative gradients are given by:

$$\boldsymbol{u}^{[j]} = \left(\begin{cases} \tau, & y_{t+h} - \hat{m}^{[j-1]}(\boldsymbol{x}_t) \geq 0 \\ \tau - 1, & y_{t+h} - \hat{m}^{[j-1]}(\boldsymbol{x}_t) < 0 \end{cases} \right)_{t=1,\ldots,T}$$

   (b) Fit each of the B base-learners specified in step 2 using the negative gradient vector $\boldsymbol{u}^{[j]}$ as the response with the corresponding input variable.

   (c) Select the best-fitting base-learner, i.e., the one that minimizes the residual sum of squares, and denote by $\hat{\boldsymbol{u}}^{[j]}$ the fitted values of the best-fitting base-learner.

(d) Update the current function estimate by adding the fitted values of the best-fitting base-learner to the function estimate of the previous iteration $j-1$:

$$\hat{\boldsymbol{m}}^{[j]} = \hat{\boldsymbol{m}}^{[j-1]} + \nu\ \hat{\boldsymbol{u}}^{[j]}$$

where $0 < \nu \leq 1$ is a shrinkage factor.

5. Stop if j has reached the maximum number of iterations J, or go to step 3.

Following the steps given above, we can see that the final function estimate $\hat{\boldsymbol{m}}$ can be written as follows:

$$\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}}^{[0]} + \sum_{j=1}^{J} \nu\ \hat{\boldsymbol{u}}^{[j]}, \tag{17}$$

and since each component $\hat{\boldsymbol{u}}^{[j]}$ depends only on one variable k, the final estimate can be written as an additive model:

$$\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}}^{[0]} + \sum_{k=1}^{p} \underbrace{\sum_{j:k \text{ is selected}} \nu\ \hat{\boldsymbol{u}}^{[j]}}_{\hat{\boldsymbol{a}}_k} \tag{18}$$

$$= \hat{\boldsymbol{m}}^{[0]} + \sum_{k=1}^{p} \hat{\boldsymbol{a}}_k, \tag{19}$$

where $\hat{\boldsymbol{a}}_k$ is the relative contribution of the variable k to the final estimate, and p is the number of initial input variables (i.e., the dimensionality of $\boldsymbol{x}_t$).

3.3 Base-learners

Several types of weak learners have been considered in the boosting literature, including regression trees (e.g., stumps, trees with two terminal nodes) (Friedman, 2001), smoothing splines (Bühlmann and Yu, 2003) and penalized regression splines (P-splines) (Schmid and Hothorn, 2008).

In this work, the base-learners with categorical variables will be estimated with standard indicator variables. For the continuous variables, we will consider one linear base-learner to model a linear effect, and a second base-learner with P-splines to model the nonlinear deviation from the linear effect, as explained in Kneib, Hothorn, and Tutz (2009).

P-splines are characterized by a number of parameters that have to be specified: the degree of the B-spline bases, the order of the difference penalty, the number of knots, and the smoothing

parameter. Cubic B-splines (i.e., of degree 3) are the most commonly used B-spline bases since they offer the best trade-off between flexibility and computational simplicity. The difference order is generally specified to be 2, i.e., deviations from linearity are penalized. Ruppert (2002) showed that the number of knots does not have much effect on the estimation provided enough knots are used.

The smoothing parameter is the main hyperparameter for P-splines; it controls the trade-off between over- and under-fitting (or equivalently, under- and over-smoothing, respectively). Specifically, the smoothing parameter is related to the weight given to the fit and penalty components in the objective function. We can parametrize the P-spline estimator in a more natural way by specifying its degree of freedom (df). Of course, for a given df, there is a corresponding smoothing parameter that can be computed. The df of the P-spline measures its "weakness", and Bühlmann and Yu (2003) and Schmid and Hothorn (2008) suggested that the df should be set to a small value (e.g., df $\in [3,4]$), and that this number should be kept fixed in each boosting iteration.

3.4 Hyperparameters selection

From the different steps described in the gradient boosting algorithm in Section 3.2, we can see that the boosting procedure depends on two hyperparameters: ν, the shrinkage factor and J, the number of boosting iterations (or equivalently, the number of boosting components). The value of ν affects the best value for J: decreasing the value of ν requires a higher value for J. Since they can both control the degree of fit, we should ideally find the best value for both of them. However, Friedman (2001) shows that small values of ν are better in that they usually avoid overfitting of the boosting procedure. Hence, there is only one hyperparameter remaining (i.e., J) for which the best value needs to be selected (Bühlmann and Yu, 2003).

After setting a range for the number of iterations, (e.g., $J \in \{1,\dots,J_{\max}\}$), the best value of the hyperparameter J can be selected in this range by time series cross-validation (also called forecast evaluation with a rolling origin (Hyndman and Athanasopoulos, 2015; Tashman, 2000)) where the first part of the time series is used as training set and the remaining part is used as validation/rolling set. In particular, the model is fitted using the training set and the errors are computed using the validation set. We repeat this procedure multiple times, and each time one example moves from the validation set to the training set. Note that we used a five-fold time series cross-validation procedure with a one-standard-error rule for model selection.

4 Electricity demand data

4.1 Smart meter data

Forecasting electricity demand is critical for electric utilities in many operational and planning tasks. Traditional electricity demand data often represent aggregated demand across many consumers (e.g., at the city or state level). In the last few years, electricity demand has been recorded at a very local level, typically for individual households, using so-called *smart meters*. These devices record and transmit electricity consumption information at 30-minute or hourly intervals, hence generating a huge quantity of data.

An important distinctive feature of smart meter data compared to traditional electricity demand data is the higher volatility due to the wider variety of individual demand patterns. With this high volatility data, it has become increasingly important to forecast not only the average electricity demand, but the entire distribution of the future demand. In other words, we need to generate probabilistic forecasts for a large set of time series.

We use the data collected during a smart metering trial conducted by the Commission for Energy Regulation (CER) in Ireland (Commission For Energy Regulation, 2011). In particular, we focus on 250 meters out of the 3639 meters associated with the residential consumers which do not have missing values. Every meter provides the electricity consumption at 30-minute intervals between 14 July 2009 and 31 December 2010; hence, each time series has 25728 observations.

A particular property of the CER data set is that it does not account for energy consumed by heating and cooling systems (Beckel et al., 2014). In fact, either the households use a different source of energy for heating, such as oil and gas, or a separate meter is used to measure the consumption due to heating. In addition, no installed cooling system has been reported in the study. The CER data set has been recently used in different studies, including Arora and Taylor (2014) and Pompey et al. (2014).

The upper time series in Figure 1 shows the electricity consumption during one week aggregated over 200 consumers belonging to the same cluster (using the CER categorization scheme), while the lower time series shows the consumption of one of the 200 consumers for the same period. We can clearly see the daily patterns for the aggregated demand, while the demand for one consumer is much more volatile and erratic, illustrating the difficulty in modeling such low-level data.

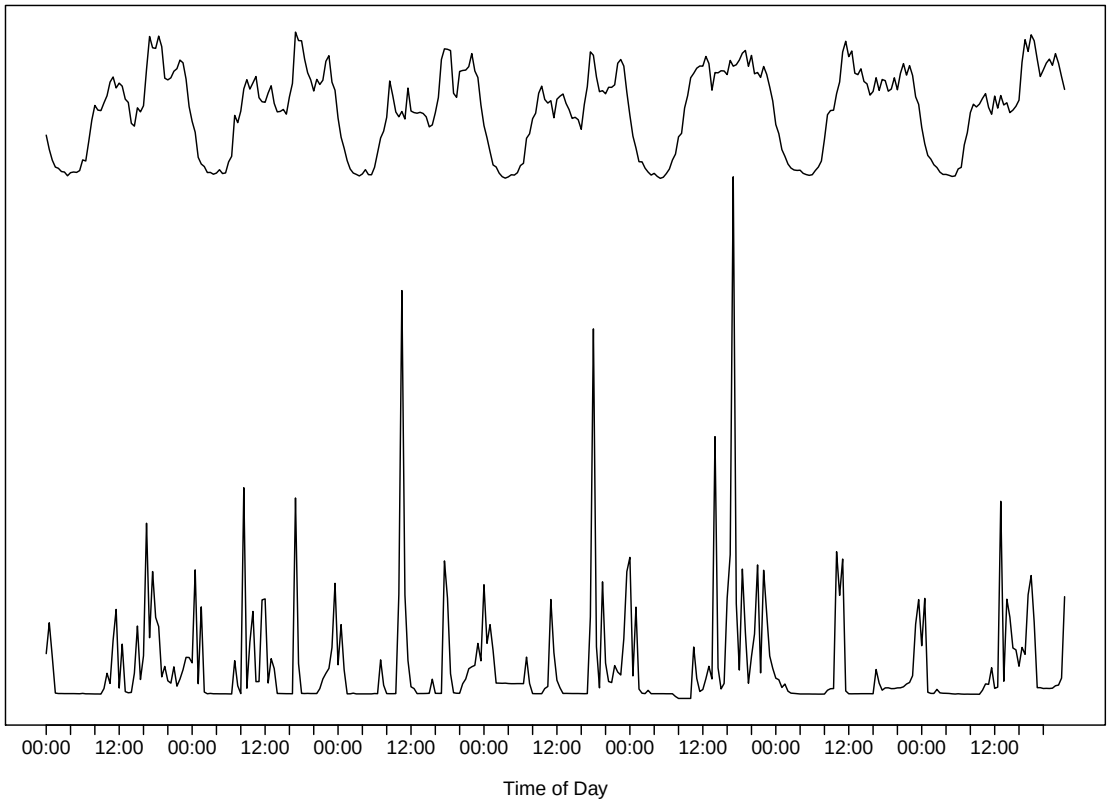

Figure 1: *Aggregated demand over 200 meters (upper time series), compared to the demand data for one of the meters (lower time series).*

4.2 Electricity demand modeling

Electricity demand is nonlinear and volatile, and is subject to a wide variety of exogenous variables, including prevailing weather conditions, calendar effect, demographic and economic variables, as well as the general randomness inherent in individual usage. How to effectively integrate the various factors into the forecasting model and provide accurate load forecasts is always a challenge for modern power industries.

Since we focus on day-ahead forecasts in this work, our models will include calendar, temperature and recent demand effects, except in model (1), where we do not include recent demand variables for $\sigma_h(\boldsymbol{x}_t)$, as suggested by Wijaya, Sinn, and Chen (2015).

Calendar effects

The calendar effects include weekly and daily seasonal patterns as well as public holidays.

- The day-of-week effect can be modeled with a factor variable, taking a different value for each day of the week.

- The time-of-day effect can be modeled with a factor variable, taking a different value for each period of the day.
- The holiday effect can be modeled with a factor variable, taking value zero on a non-work day, some non-zero value on the day before a non-work day and a different value on the day after a non-work day.

Calendar variables are important predictors for electricity demand. For example, we expect a lower demand with a lower uncertainty during the night, and a larger demand and higher uncertainty during the day.

Temperature effects

Due to thermal inertia in buildings, it is important to consider lagged temperatures as well as current temperatures in any demand forecasting model. The temperature effects include

- The current temperature and temperatures from the preceding 12 half-hours, and for the equivalent half-hour on each of the previous two days;
- The minimum and maximum temperature in the last 24 hours;
- The average temperature for the last seven days.

Because the smart meters in our data set do not account for heating and cooling, the temperature effect is expected to be small. Nevertheless, cold weather tends to be associated with higher usage, simply because of greater indoor activity.

Recent demand effects

We incorporate recent demand values into the model as follows:

- Lagged demand for each of the preceding (available) 12 half-hours, and for the equivalent half-hour in each of the previous two days.
- The minimum and maximum demand in the last 24 hours.
- The average demand for the last seven days.

By doing this, the serial correlations within the demand time series can be captured within the model, and the variations of demand level over time can be embedded in the model as well.

5 Experiments

5.1 Setup and preprocessing

We focus on day-ahead probabilistic forecasts using the smart meter dataset described in Section 4.1. In other words, we forecast the electricity demand for each hour for the next 24 hours once per day at a specific time of day. In the following, we present the experimental setup and preprocessing; in Section 5.2 we list the forecasting methods that we compared; in Section 5.3, we present and discuss the results.

For each of the 250 smart meters, we have access to a time series representing half-hourly electricity consumption for almost 18 months, more precisely $T = 25728$ observations.

We used the first 12 months as training period to fit the different models, and the remaining data as a test set to evaluate forecast accuracy. The last month of the training period is used for cross-validation. The testing period is used to generate forecasts for lead-times ranging from one-step to $H = 48$-steps ahead with each observation as a forecast origin. We measure forecast accuracy using the CRPS and the unconditional coverage, both defined in Section 2.4. In expressions (11) and (13), we have $N \approx 7300$ (5 months of half-hourly data) and we use $M = 10^5$ random samples to approximate the expectations.

Since temperature data are not provided, and the location of each meter is anonymized for confidentiality reasons, we downloaded half-hourly weather data for the Dublin airport from `wunderground.com`. A similar approach has been considered in (Pompey et al., 2014). We make the assumption that Ireland is sufficiently small so that the weather at Dublin airport is similar (i.e., at least positively correlated) to the weather elsewhere in the country at any given time.

In order to allow averaging of accuracy metrics across different meters, consumption observations are divided by their maximum value. Also, since with disaggregated data there are many observations close to zero, we have applied a square root transformation to guarantee the non-negativity of the final forecasts; the square root is also known to stabilize the variance when the latter is proportional to the mean (as, e.g., for the Poisson distribution). No transformation has been applied to the aggregated smart meter data.

In the gradient boosting procedure, we use a degree of freedom df = 4 for the P-splines. The shrinkage factor is set to $\nu = 0.3$, and is kept fixed for all iterations. Finally, the number of boosting iterations J is selected by cross-validation in the set $\{1, \ldots, J_{\max}\}$, where $J_{\max} = 200$ gives

a good tradeoff between accuracy and computational time. Our implementation of the gradient boosting procedure is based on the *mboost* package (Hothorn et al., 2010) available for the R programming language (R Core Team, 2015).

5.2 Forecasting methods

1) Conditional mean and variance forecasting with boosted additive models. This is the method presented in Section 2.2, where a regression model is estimated for both the conditional mean and variance, and a Gaussian assumption is made to obtain the full conditional distribution. This approach has been considered in Wijaya, Sinn, and Chen (2015) to generate probabilistic forecasts for aggregated electricity demand using backfitted additive models. Notice that fitting a normal distribution after applying a square-root transformation to the data is equivalent to fitting a chi-squared distribution with one degree of freedom to the initial data.

We are considering a variant of this approach, where the models are fitted using boosted additive models (see Section 3). This method will be denoted `NORMAL-GAMBOOST` and abbreviated as `NORMAL`.

2) Quantile forecasting with boosted additive models. This is the method presented in Section 2.3, where a quantile regression model is estimated for each τ-quantile of the distribution, with $\tau = 0.01, 0.02, \ldots, 0.99$; no distributional assumption is made here. As with the previous method, each model is fitted using boosted additive models. This method will be denoted `QR-GAMBOOST` and abbreviated as `QR`.

3) Quantiles computed conditional on the period of the day This method segments the data by period of the day, into 48 sub-datasets, and the different quantiles are computed for each sub-dataset. With this benchmark method, we allow the distribution to change for each period of day, but temperatures or lagged demands are not accounted for. A variant of this benchmark method has been used in Arora and Taylor (2014) with kernel density estimation methods. This method will be denoted `PeriodOfDay`.

4) Unconditional quantiles. We compute the τ-quantile of the distribution of all the historical observations. In other words, this method does not condition on recent demand or temperature observations, and it does not attempt to capture the seasonality in the smart meter data. This method will be denoted `Uncond`.

5.3 Results

The first panel of Figure 2 shows the CRPS defined in (11) averaged over all meters for all the forecasting methods over the forecast horizon. The right and bottom panels show the CRPS as given in (13), decomposed into the reliability and spread components, respectively. As with the traditional bias and variance decomposition, it is important to consider both terms when comparing forecasting methods. Figure 3 gives the unconditional coverage for the different methods by plotting the points $(\tau, \tilde{\tau}^{(1)})$ and $(\tau, \tilde{\tau})$ for $\tau = 0.01, 0.02, \ldots, 0.99$, where $\tilde{\tau}^{(1)}$ and $\tilde{\tau}$ have been defined in (14) and (15), respectively. Finally, Figure 4 shows an example of density forecasts for the different forecasting methods.

In the top left panel of Figure 2, we can see that `Uncond` has the worst performance. By conditioning on the period of the day, `PeriodOfDay` significantly improves the results, confirming the importance of having calendar variables as predictors for electricity demand forecasting. The remaining panels of Figure 2 show that `PeriodOfDay` has achieved a lower CRPS than `Uncond` by reducing both reliability and spread. In Figure 4, we can see that the predictive distributions

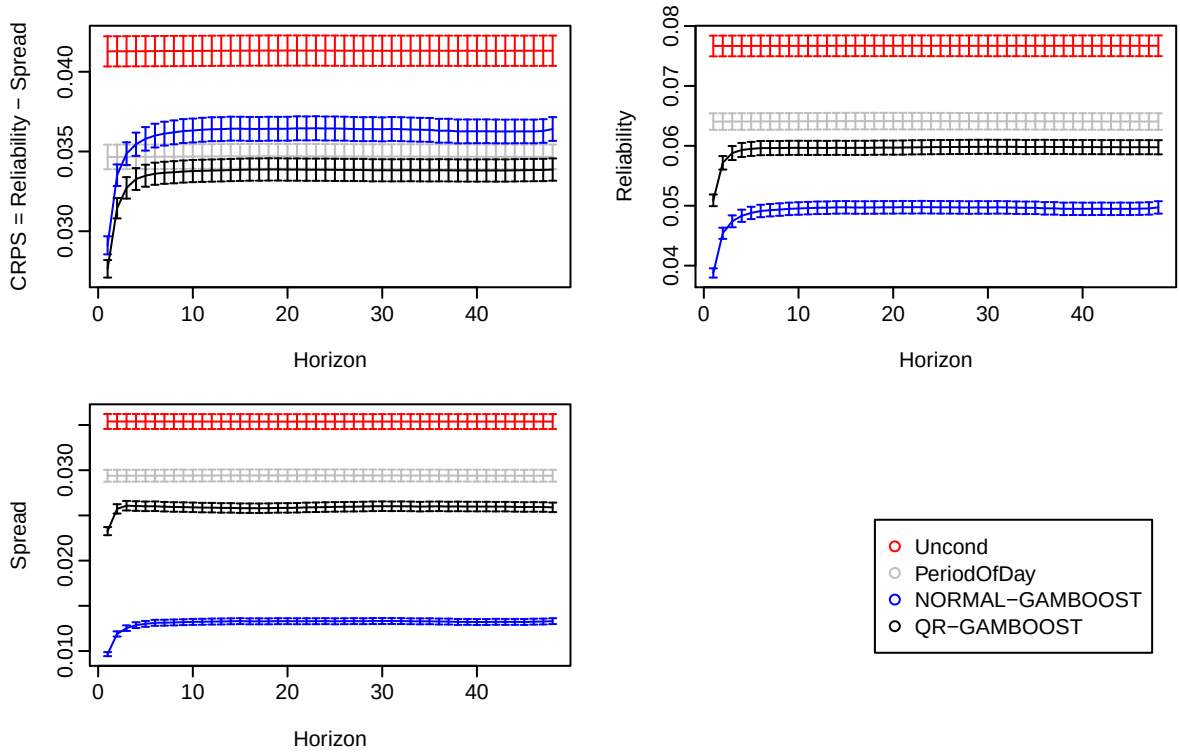

Figure 2: *The CRPS averaged over all meters for the different methods over the forecast horizon decomposed into reliability and spread. The error bars give the standard errors.*

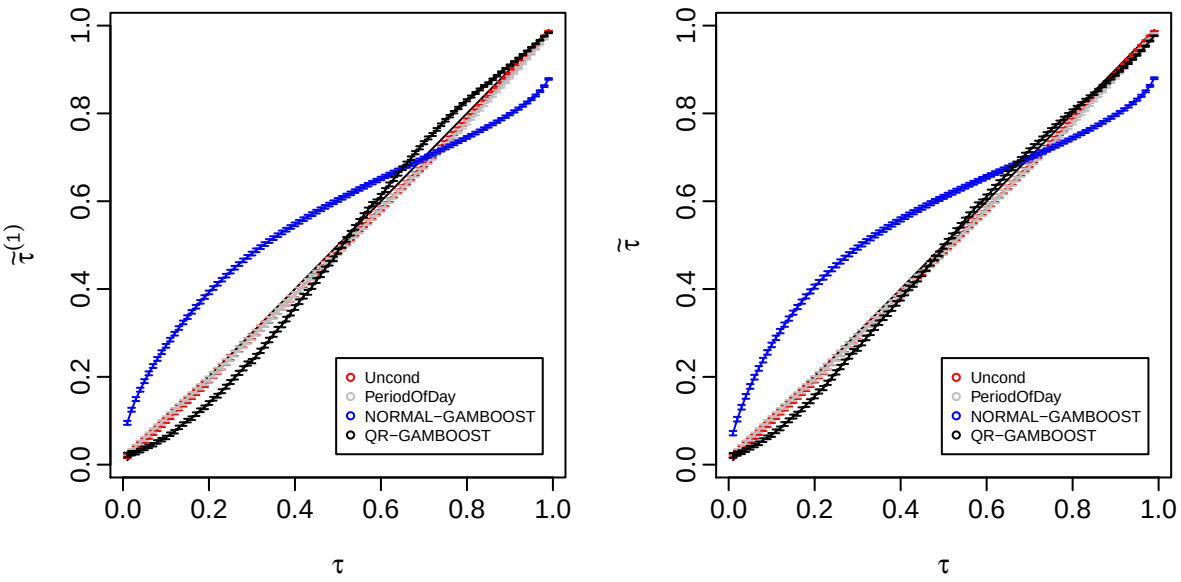

Figure 3: *The unconditional coverage averaged over all meters for horizon $h = 1$ (left), and averaged over all forecast horizons (right). The error bars give the standard errors.*

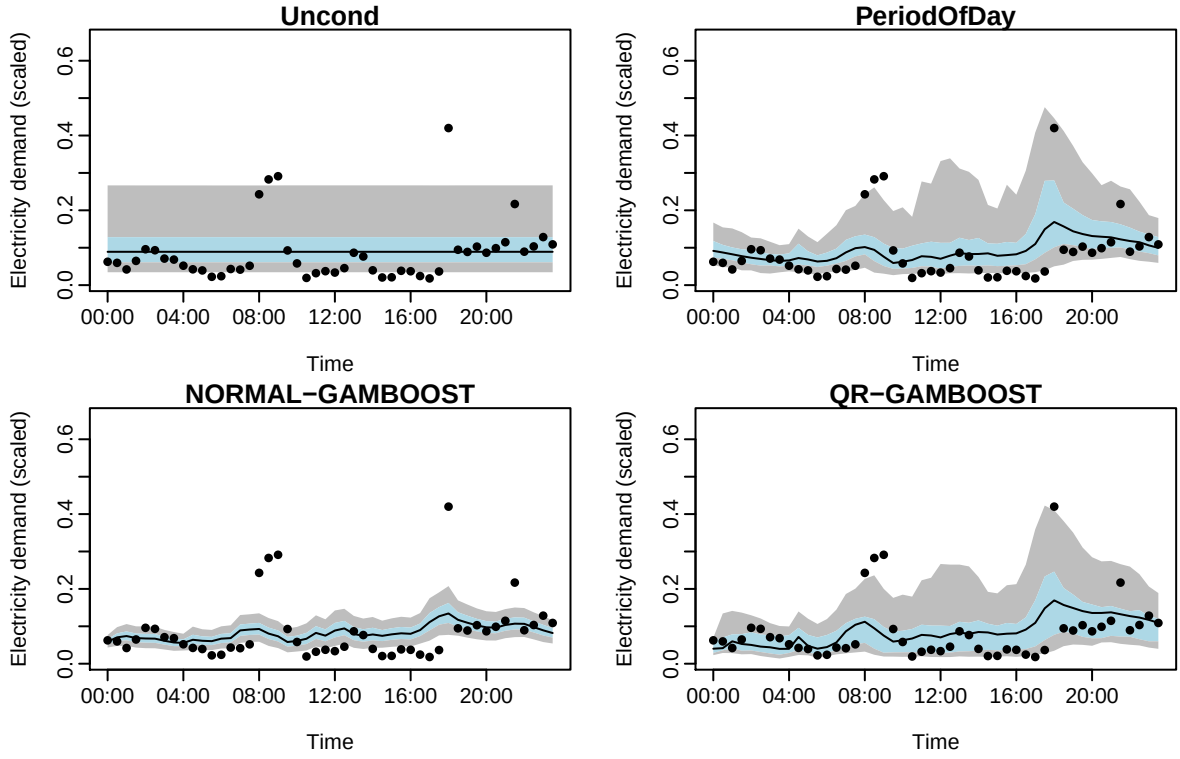

Figure 4: *One-day ahead density forecasts of the four forecasting methods for the meter 2. The blue and grey regions are 50% and 90% prediction intervals, respectively.*

of PeriodOfDay have a spread changing with the forecast horizon (or equivalently with the period of day), while Uncond has a constant spread.

Figure 2 also reveals that, for QR and NORMAL, the CRPS at the first few horizons is particularly small compared to Uncond and PeriodOfDay. This is because the recent demand (used by the two methods) is a good predictor for the first few horizons. Also, after the first few horizons, QR has a CRPS close to that of PeriodOfDay. This can be explained by the fact that after the first few horizons, the recent demand is no longer a good predictor, but the calendar variables become the main predictors.

We can also see that QR outperforms NORMAL, which suggests that the normality assumption (after applying a square-root transformation on the demand) is not valid for individual electricity consumption. This is also confirmed in Figure 3 where we can see that NORMAL has the worst unconditional coverage compared to the other methods.

In the right and bottom panels of Figure 2, we can see that NORMAL has both a better reliability and spread than QR. However, the fact that it has a higher CRPS than QR indicates that the predictive distributions of NORMAL are not sufficiently spread out to match the true uncertainty of the demand. Although QR has both higher reliability and spread, the lower CRPS suggests that it generates predictive densities that are closer to the true densities. By comparing the predictive densities of NORMAL and QR in Figure 4, we can see that QR better matches the realized observations than NORMAL. The previous results confirm the advantage of considering quantile regression methods to benefit from a higher flexibility for the predictive densities when forecasting individual electricity consumption.

Let us now consider the aggregated demand obtained by summing electricity demand over all the meters. Recall that aggregated demand is less volatile than individual consumption data, as shown in Figure 1. Figures 5, 6 and 7 give the same information as in Figures 2, 3 and 4 but for the aggregated demand.

In contrast to the results obtained for the disaggregated demand, we can see in Figure 5 that NORMAL has a lower CRPS than QR. This can be explained by the fact that the more meters we aggregate the more normally distributed is the aggregated demand, as a consequence of the Central Limit Theorem. The normal distribution for aggregated demand has also been observed in Sevlian, Patel, and Rajagopal (2014), while Sevlian and Rajagopal (2013) and Sevlian and Rajagopal (2014) have studied the effect of aggregation on short-term average electricity demand forecasts.

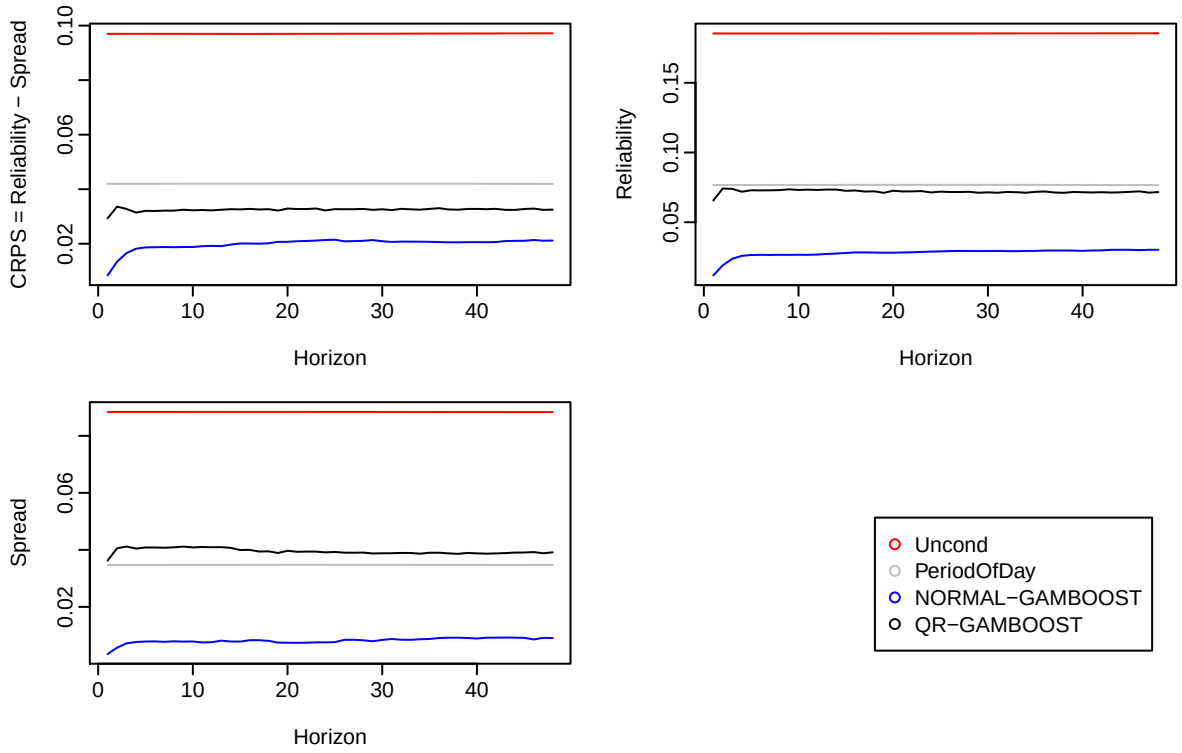

Figure 5: *The CRPS of the different methods for the aggregated demand over the forecast horizon decomposed into reliability and spread.*

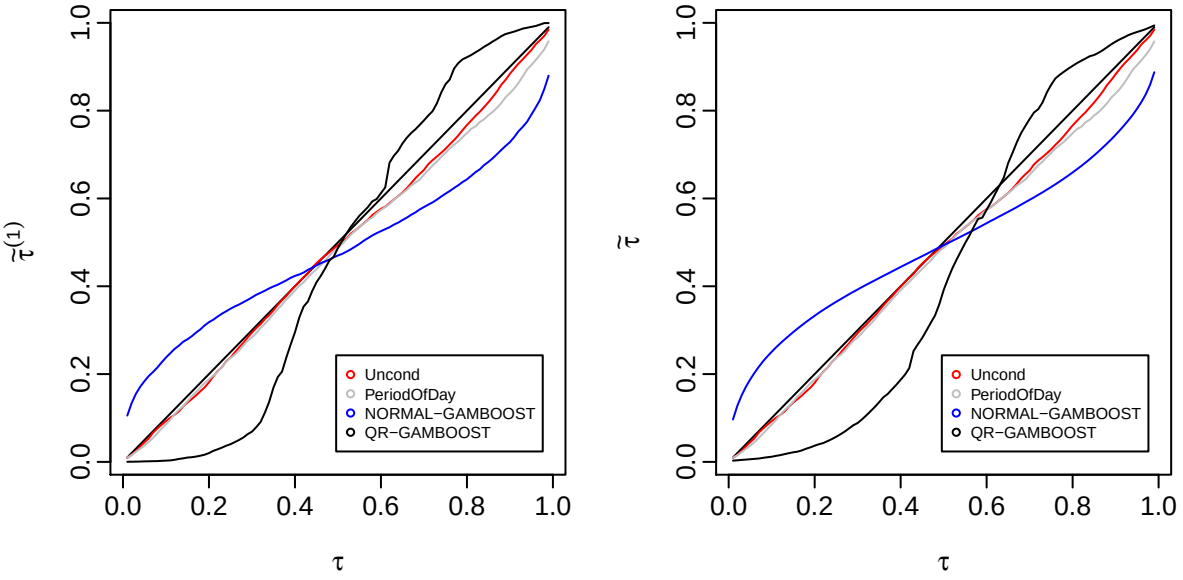

Figure 6: *The unconditional coverage averaged for the aggregated demand at horizon $h = 1$ (left), and averaged over all forecast horizons (right).*

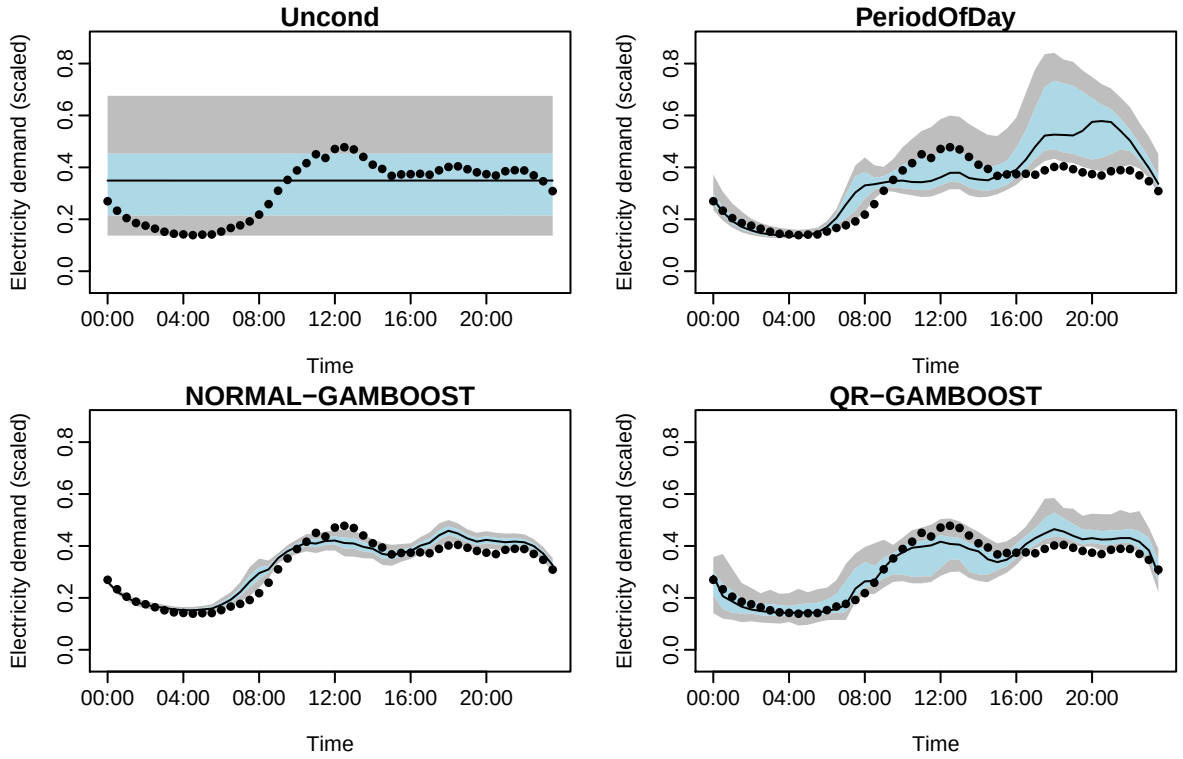

Figure 7: *One-day ahead density forecasts of the four forecasting methods for the aggregated demand. The blue and grey regions are 50% and 90% prediction intervals, respectively.*

The bottom panel of Figure 5 shows that the predictive densities of NORMAL are relatively sharp compared to other methods, but at the same time provide a better CRPS, as can be seen in the top left panel. In Figure 7, we can also see that the sharper density forecasts of NORMAL allow to obtain a relatively good coverage since the demand is much smoother than individual electricity consumption, as illustrated in Figure 4.

6 Conclusions and future work

Probabilistic forecasting is more challenging than point forecasting since we need to forecast not only the conditional mean but the entire conditional distribution of the future observations.

We have presented two different methods to generate probabilistic forecasts: the first method is based on traditional regression and involves forecasting the conditional mean and variance of the future observations; this allows to obtain point forecasts and to provide some information spread of the future observations around the mean. The second approach is based on quantile regression and involves forecasting a set of quantiles of the future distributions.

We proposed an implementation for the two methods based on boosted additive models, which enjoy many useful features including accuracy, flexibility, interpretability and automatic variable selection.

Generating accurate probabilistic time series forecasts is particularly relevant in many energy applications. We have considered the problem of probabilistic forecasting for electricity smart meter data. The results of the comparison between the two methods can be summarized as follows.

At the disaggregated level, with the large diversity of consumption behavior and the high volatility of the demand, we found that quantile forecasts outperform forecasts based on a normal distribution. The decomposition of the forecast errors shows that normal forecasts produce predictive densities, which are too concentrated, not matching the true uncertainty.

At the aggregated level, where the demand becomes more normally distributed as a consequence of the Central Limit Theorem, normal forecasts provide better forecasts than quantile forecasts. The decomposition of the forecast errors shows that the quantile forecasts lack sharpness, that is the forecasts are more spread out than necessary to match the true uncertainty.

These results are particularly useful since a large body of literature has so far focused on forecasting the electricity demand at the aggregated level, while more data is becoming available at the disaggregated level.

For future work, we will investigate the problem of forecasting the peak electricity demand, that is quantile forecasts for $\tau > 0.99$, both at the disaggregated and aggregated level. Another important direction is to improve the computational load of probabilistic forecast methods since, in practice, utilities need to deal with millions of smart meters.

7 Acknowledgments

The authors thank Prof. Pierre Pinson for helpful discussions.

References

Alfares, HK and M Nazeeruddin (2002). Electric load forecasting: Literature survey and classification of methods. *International journal of systems science* **33**(1), 23–34.

Arora, S and JW Taylor (2014). Forecasting Electricity Smart Meter Data Using Conditional Kernel Density Estimation. *Omega*.

Bacher, P, H Madsen, and HA Nielsen (2009). Online short-term solar power forecasting. *Solar Energy* **83**(10), 1772–1783.

Beckel, C, L Sadamori, T Staake, and S Santini (2014). Revealing household characteristics from smart meter data. *Energy* **78**, 397–410.

Ben Taieb, S and RJ Hyndman (2014). A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting* **30**(2), 382–394.

Bühlmann, P and B Yu (2003). Boosting With the L2 Loss: Regression and Classification. *Journal of the American Statistical Association* **98**(462), 324–339.

Chen, LH, MY Cheng, and L Peng (2009). Conditional variance estimation in heteroscedastic regression models. *Journal of Statistical Planning and Inference* **139**(2), 236–245.

Chernozhukov, V, I Fernández-Val, and A Galichon (2010). Quantile and probability curves without crossing. *Econometrica: Journal of the Econometric Society* **78**(3), 1093–1125.

Cho, H, Y Goude, X Brossat, and Q Yao (2013). Modeling and Forecasting Daily Electricity Load Curves: A Hybrid Approach. *Journal of the American Statistical Association* **108**(501), 7–21.

Commission For Energy Regulation (2011). *Electricity smart metering customer behaviour trials findings report*. Tech. rep. Dublin: Commission for Energy Regulation.

Engle, RF, CWJ Granger, R Ramanathan, and F Vahid-Araghi (1993). *Probabilistic Methods in Forecasting Hourly Loads*. Tech. rep. TR-101902. Electric Power Research Institute. http://www.epri.com/abstracts/Pages/ProductAbstract.aspx?ProductId=TR-101902.

Fan, J and Q Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**(3), 645–660.

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

Friedman, J and T Hastie (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Annals of Statistics* **28**(2), 337–407.

Friedman, JH (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics* **29**(5), 1189–1232.

Gneiting, T (2011). Quantiles as optimal point forecasts. *International Journal of forecasting* **27**(2), 197–207.

Gneiting, T, F Balabdaoui, and AE Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **69**(2), 243–268.

Gneiting, T and M Katzfuss (2014). Probabilistic Forecasting. *Annual Review of Statistics and Its Application* **1**(1), 125–151.

Gneiting, T and AE Raftery (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102**(477), 359–378.

Groen, JJJ, R Paap, and F Ravazzolo (2013). Real-Time Inflation Forecasting in a Changing World. *Journal of business & economic statistics: a publication of the American Statistical Association* **31**(1), 29–44.

Hastie, TJJ and RJJ Tibshirani (1990). *Generalized additive models*. Chapman & Hall/CRC.

Hastie, TJ, R Tibshirani, and JH Friedman (2008). *The elements of statistical learning*. Vol. 18. Springer-Verlag, p. 764.

Hippert, HS, CE Pedreira, and RC Souza (2001). Neural networks for short-term load forecasting: a review and evaluation. *IEEE Transactions on Power Systems* **16**(1), 44–55.

Hong, T (2010). "Short Term Electric Load Forecasting". PhD thesis.

Hothorn, T, P Bühlmann, T Kneib, M Schmid, and B Hofner (2010). Model-based boosting 2.0. *Journal of Machine Learning Research: JMLR* **11**, 2109–2113.

Hyndman, RJ and G Athanasopoulos (2015). *Forecasting: principles and practice*. Melbourne, Australia: OTexts.

Jones, HE and DJ Spiegelhalter (2012). Improved probabilistic prediction of healthcare performance indicators using bidirectional smoothing models. *Journal of the Royal Statistical Society. Series A* **175**(3), 729–747.

Kneib, T (2013). Beyond mean regression. *Statistical Modelling* **13**(4), 275–303.

Kneib, T, T Hothorn, and G Tutz (2009). Variable Selection and Model Choice in Geoadditive Regression Models. *Biometrics* **65**(003), 626–634.

Koenker, R (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Lou, Y, R Caruana, J Gehrke, and G Hooker (2013). Accurate intelligible models with pairwise interactions. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp.623–631.

Mayr, A, T Hothorn, and N Fenske (2012). Prediction intervals for future BMI values of individual children: a non-parametric approach by quantile boosting. *BMC Medical Research Methodology* **12**, 6.

Palmer, TN (2012). Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society* **138**(665), 841–861.

Pompey, P, A Bondu, Y Goude, and M Sinn (2014). "Massive-Scale Simulation of Electrical Load in Smart Grids using Generalized Additive Models". In: *Lecture Notes in Statistics: Modeling and Stochastic Learning for Forecasting in High Dimension*. Springer.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Ruppert, D (2002). Selecting the Number of Knots for Penalized Splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

Schapire, RE (1990). The strength of weak learnability. *Machine Learning* **5**(2), 197–227.

Schapire, RE and Y Freund (2012). *Boosting: Foundations and Algorithms*. The MIT Press.

Schmid, M and T Hothorn (2008). Boosting Additive Models using component-wise P-Splines. *Computational Statistics & Data Analysis* **53**(002), 298–311.

Sevlian, R and R Rajagopal (2013). Value of aggregation in smart grids. In: *Smart Grid Communications (SmartGridComm), 2013 IEEE International Conference on*, pp.714–719.

Sevlian, R, S Patel, and R Rajagopal (2014). Distribution System Load and Forecast Model. arXiv: 1407.3322 [stat.AP].

Sevlian, R and R Rajagopal (2014). A Model For The Effect of Aggregation on Short Term Load Forecasting. In: *IEEE Power and Energy Society General Meeting*.

Spiegelhalter, DJ (2014). Statistics. The future lies in uncertainty. *Science* **345**(6194), 264–265.

Tao, H and S Fan (2014). "Probabilistic Electric Load Forecasting: A Tutorial Review". Submitted to International Journal of Forecasting.

Tashman, LJ (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* **16**(4), 437–450.

Taylor, JW (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research* **204**(1), 139–152.

Wijaya, TK, M Sinn, and B Chen (2015). Forecasting Uncertainty in Electricity Demand. In: *AAAI-15 Workshop on Computational Sustainability*.

Zheng, J, DW Gao, and L Lin (2013). Smart Meters in Smart Grid: An Overview. In: *Green Technologies Conference, 2013 IEEE*, pp.57–64.

Zhu, X and MG Genton (2012). Short-Term Wind Speed Forecasting for Power System Operations. *International Statistical Review* **80**(1), 2–23.