Forecasting Uncertainty in Electricity Smart Meter Data by Boosting Additive Quantile Regression

Souhaib BenTaieb, Raphaël Huser, Rob J. Hyndman and Marc G. Genton

***Abstract*—Smart electricity meters are currently deployed in millions of households to collect detailed individual electricity consumption data. Compared to traditional electricity data based on aggregated consumption, smart meter data are much more volatile and less predictable. There is a need within the energy industry for probabilistic forecasts of household electricity consumption to quantify the uncertainty of future electricity demand, in order to undertake appropriate planning of generation and distribution. We propose to estimate an additive quantile regression model for a set of quantiles of the future distribution using a boosting procedure. By doing so, we can benefit from flexible and interpretable models which include an automatic variable selection. We compare our approach with three benchmark methods on both aggregated and disaggregated scales using a smart meter dataset collected from 3639 households in Ireland at 30-minute intervals over a period of 1.5 years. The empirical results demonstrate that our approach based on quantile regression provides better forecast accuracy for disaggregated demand while the traditional approach based on a normality assumption (possibly after an appropriate Box-Cox transformation) is a better approximation for aggregated demand. These results are particularly useful since more energy data will become available at the disaggregated level in the future.**

I. Introduction

THE energy sector has been changing dramatically, notably due to the integration of renewable energy sources, in 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 for utilities to quantify the uncertainty in future generation and demand, especially for wind power [1], solar power [2] and electricity demand [3]. In particular, accurate probabilistic forecasts for electricity demand are 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; i.e. in forecasting the mean of the future demand distribution. These forecasts are often conditional on a number of predictor variables such as calendar and temperature variables. Many models have been considered for modeling and forecasting the average electricity load; these are comprehensively reviewed in [4]. The literature on probabilistic load forecasting is much more sparse and is reviewed in [5].

In this article, we focus on the problem of probabilistic forecasting for smart meter data. A *smart meter* is an electronic device that records and transmits electricity consumption information typically at intervals from 1 minute to one hour, hence generating a huge quantity of data. Electric load forecasts at the household level can be particularly useful for evaluating demand response programs [6] as well as for improving forecasts at aggregated levels. Compared to traditional electricity load, smart meters measure the load at a very local level, typically for individual households. Smart meter data are highly volatile, and 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, we need to move from point forecasting to probabilistic forecasting [7].

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 [6] who considered kernel density estimation methods to generate probabilistic forecasts for individual smart meters. Other papers on smart meter forecasting, e.g. [8], [9], have mainly focused on point forecasting. One of the contributions we make in this paper is to enrich the literature on probabilistic forecasting for electricity smart meter data.

Probabilistic forecasting methods can be classified into parametric and non-parametric approaches [10]. With parametric approaches, one assumes that the predictive distribution follows some known parametric distribution (e.g. a normal distribution), and the associated parameters need to be estimated from the data. See [11] and [3] for examples. On the other hand, non-parametric approaches estimate the predictive distribution with only weak assumptions (such as smoothness) on the shape of the distribution. Examples of non-parametric approaches include kernel density estimation [12] and quantile regression [13].

In this article, we propose to generate probabilistic forecasts for electricity smart meter data using quantile regression, where a different forecasting model is estimated for each quantile of the distribution of the future demand. In particular, we make the following contributions:

1) We show how to move from average electric load forecasting (Section III) to probabilistic load forecasting by quantile regression (Section IV).

S. BenTaieb is with the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, and the Monash Business School, Clayton, VIC 3800, Australia (e-mails: souhaib.bentaieb@kaust.edu.sa, souhaib.bentaieb@monash.edu).

R. Huser is with the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia (e-mail: raphael.huser@kaust.edu.sa).

R. J. Hyndman is with the Monash Business School, Clayton, VIC 3800, Australia (e-mail: rob.hyndman@monash.edu).

M. G. Genton is with the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia (e-mail: marc.genton@kaust.edu.sa).

2) We propose a probabilistic forecasting method based on boosting additive quantile regression that allows flexibility, interpretability and automatic variable selection (Section VI).
3) We compare the proposed method with three other benchmark methods using 3639 meters (with more than 25000 observations for each meter) from a public smart meter data set (Section VII). The methods are compared on disaggregated and aggregated electricity demand using the continuous ranked probability score (CRPS).

II. Smart meter data

We use the data collected during a smart metering trial conducted by the Commission for Energy Regulation (CER) in Ireland [14]. The data set contains measurements of electricity consumption gathered from 4225 residential consumers and 2210 small to medium-sized enterprises (SMEs). We focus on 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. As claimed by [15], to the best of their knowledge, the CER data set does not account for energy consumed by heating and cooling systems. 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 [6], [9], [15], [16].

A distinctive feature of smart meter data compared to traditional electric load data is the higher volatility due to the variety of individual demand patterns. In fact, traditional electric load data often represent the aggregated load across many consumers (e.g. at the city level), whereas smart meters measure the load at a very local level, typically for individual households, where usage behavior can vary greatly.

The upper time series in Figure 1 shows the electricity consumption during one week aggregated over 1000 consumers belonging to the same cluster (using the CER categorization scheme), while the lower time series shows the consumption of one of the 1000 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.

Electricity demand is subject to a wide variety of exogenous variables, including calendar effects and weather conditions. For example, the period of day and the day of week are important predictors for electricity demand, since we expect a lower demand with a lower uncertainty during the night, and a larger demand and higher uncertainty during the day.

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.

Smart meter data will also typically exhibit serial dependence within the demand time series, although it is expected to be smaller than with aggregated demand.

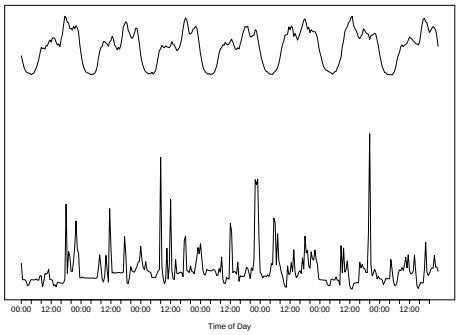

Fig. 1. Aggregated demand over 1000 meters (upper time series), compared to the demand data for one of the meters (lower time series).

III. Average electric load forecasting

In order to generate accurate load forecasts, the forecasting model should effectively combine the various predictors. We first describe how to model the average electricity load to generate point forecasts, while in the next section we show how to move from point forecasts to probabilistic forecasts using quantile regression.

We can model the average electric load at time $t+h$ using a different model at each forecast horizon h as follows:

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

where $\boldsymbol{x}_t = (\boldsymbol{y}_t, \boldsymbol{z}_{t+h})$,

- y_t denotes the demand at time t;
- $\boldsymbol{z}_t$ is a vector of exogenous variables known at time t;
- $\boldsymbol{y}_t$ is a vector of past demands occurring prior to time t;
- ε_t denotes the model error with $\mathbb{E}[\varepsilon_t] = 0$ and $\mathbb{E}[\boldsymbol{x}_t \varepsilon_{t+h}] = 0$.

In order to model the calendar effects on demand, we included multiple calendar variables as exogenous variables, namely the period-of-day, day-of-week, time-of-year and holiday variables. To model the effects of recent temperatures on the demand, we also included current and lagged temperature variables as exogenous variables. Finally, the vector $\boldsymbol{y}_t$ incorporates recent demand values including the lagged demand for each of the preceding 12 half-hours, and for the equivalent half-hours in each of the previous two days. This allow us to capture the serial correlations within the demand time series as well as the variations of demand level throughout the time.

The assumption on the error term ε_t in expression (1) implies that the predictors describe the expectation of the response y_{t+h}. In addition, if the functions g_h are estimated using squared errors, then the problem reduces to the estimation of the conditional expectation $\mathbb{E}[y_{t+h} \mid \boldsymbol{x}_t]$. In that case, we are applying mean regression, and the associated point forecasts are conditional means.

IV. Probabilistic load forecasting by quantile regression

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 forecast horizon h and forecast time origin t, where y_t and $\boldsymbol{x}_t$ are defined as above.

A popular approach to generate probabilistic forecasts is to make assumptions about the form of the conditional distribution (e.g. a normal distribution), and to estimate the corresponding parameters from data. See [3] for electricity demand, and [11], [17] and [10] for wind energy.

A more general approach than making assumptions about the form of the conditional distribution consists in computing 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. 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

$$y_{t+h} = g_{h,\tau}(\boldsymbol{x}_t) + \varepsilon_{t+h,\tau}, \quad F_{\varepsilon_{t+h},\tau}(0) = \tau, \tag{2}$$

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:

$$q_{t,h}^{(\tau)} = F_{t,h}^{-1}(\tau \mid \boldsymbol{x}_t) = g_{h,\tau}(\boldsymbol{x}_t). \tag{3}$$

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, the conditional τ-quantile $q_{t,h}^{(\tau)}$ can be computed using the pinball loss function [18]:

$$q_{t,h}^{(\tau)} = \arg\min_{q} \mathbb{E}[L_\tau(Y, q) \mid \boldsymbol{x}_t], \tag{4}$$

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

$$L_\tau(y, q) = \begin{cases} \tau(y - q) & \text{if } y \geq q; \\ -(1 - \tau)(y - q) & \text{if } y < q. \end{cases} \tag{5}$$

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 (4) 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:

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

However, since each τ-quantile is modeled and estimated independently for each probability τ, the monotonicity property might be not satisfied for all quantiles; this problem is known as *quantile crossing* [13]. Multiple approaches have been proposed to deal with the problem of quantile crossing, including joint estimation or monotone rearranging [19]; the latter is the approach that we adopt in this work.

In the energy context, quantile regression has been notably considered to forecast wind energy [20] and electricity prices [21]. However, except for [22], it has received little attention in the load forecasting literature.

V. 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), defined as follows [23]:

$$\text{CRPS}(\hat{F}_{t,h}, y_{t+h}) = \int_{-\infty}^{\infty} \Big(\hat{F}_{t,h}(z) - \mathbb{1}(z \geq y_{t+h}) \Big)^2 \, dz \tag{6}$$

where $\mathbb{1}$ is the indicator function.

Compared to other forecast accuracy measures such as the probability integral transform (PIT), the CRPS quantifies both the *calibration* and *sharpness* of the probabilistic forecast [24]. Calibration, also called *reliability*, 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 quantiles. The sharpness measures the *concentration* of the predictive distributions and is a property of the forecasts only. The best forecast is the one that maximizes the sharpness of the predictive distributions subject to calibration [24].

The CRPS can be defined equivalently as follows:

$$\text{CRPS}(\hat{F}_{t,h}, y_{t+h}) = 2 \int_0^1 L_\tau\Big(y_{t+h}, \hat{F}_{t,h}^{-1}(\tau)\Big) \, d\tau \tag{7}$$

$$= \underbrace{\mathbb{E}_{\hat{F}_{t,h}} |Y - y_{t+h}|}_{\text{Absolute differences}} - \underbrace{\frac{1}{2} \mathbb{E}_{\hat{F}_{t,h}} |Y - Y'|}_{\text{Spread}}, \tag{8}$$

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

In this work, we use the definition given in expression (8) since it allows us to decompose the CRPS into the *absolute differences* and *spread* components. The former corresponds to the average absolute difference between realised observations and values sampled from $\hat{F}_{t,h}$, and reduces to the absolute error if $\hat{F}_{t,h}$ is a point forecast. The latter measures the lack of sharpness of the predictive distribution with cumulative distribution function (CDF) $\hat{F}_{t,h}$.

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

$$\text{CRPS}(h) = \frac{1}{N} \sum_{t=1}^{N} \text{CRPS}(\hat{F}_{t,h}, y_{t+h}). \tag{9}$$

In particular, if we use the definition of the CRPS in expression (8), we can approximate the CRPS at horizon h as follows:

$\text{CRPS}(h)$

$$= \underbrace{\left(\frac{1}{NM}\sum_{t=1}^{N}\sum_{i=1}^{M}|y_{t+h}^{(i)} - y_{t+h}|\right)}_{\text{Estimated absolute differences}} - \underbrace{\left(\frac{1}{2NM}\sum_{t=1}^{N}\sum_{i=1}^{M}|y_{t+h}^{(i)} - y_{t+h}^{(i)'}|\right)}_{\text{Estimated spread}}, \quad (10)$$

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 $\hat{F}_{t,h}$.

Finally, 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 (5).

VI. Boosting additive quantile regression

There are a range of quantile regression models available. In this work, we will consider additive quantile regression models estimated by gradient boosting [25] since they allow for more flexibility than linear models, more interpretability than nonlinear blackbox models, as well as an automatic variable selection. Furthermore, additive models have been successfully applied in many studies for average electric load forecasting [26], [27].

A. Additive quantile regression

Under the additivity assumption, the conditional τ-quantile at forecast horizon h in (3) can be expressed as

$$g_{h,\tau}(\boldsymbol{x}_t) = a_0 + a_1(x_{1t}) + \cdots + a_p(x_{pt}), \quad (11)$$

where $a_k(\cdot)$ 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 additive models is their flexibility: since each function a_k can possibly have a non-linear shape, additive models 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, additive models 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 additive models 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 practice, however, the additivity assumption is not too strong since higher-order interactions are often weak. Moreover, our approach can easily include interactions terms, e.g. by specifying bivariate learners [28].

Backfitting [29] and gradient boosting [30] are the two popular methods for fitting additive models. [31] has shown that the boosting procedure is competitive with respect to backfitting and can even outperform the latter in high dimensions p. Estimation of additive models based on total variation penalties have also been proposed in [32] and [33]. In this work, we will consider a gradient boosting procedure to estimate additive quantile regression models.

B. 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", i.e. with high bias and small variance. Since its inception, boosting has attracted much attention due to its excellent prediction performance in a wide range of applications [34]. 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 [30].

Given a dataset $\mathcal{D} = \{(y_{t+h}, \boldsymbol{x}_t)\}_{t=1}^{T}$ where y_t and $\boldsymbol{x}_t$ are linked through (11), and a loss function $L(y, m)$, the goal is to fit the model (11) by minimizing the loss L over the dataset $\mathcal{D}$. In the following, we will present the gradient boosting procedure for quantile regression (also called quantile boosting [25]), which corresponds to using L_τ defined in (5) as the loss function. A similar procedure can be used for mean regression by using the $L_2(y, m) = (y - m)^2$ loss function (also called L_2Boost [31]).

Denote by $\hat{\boldsymbol{m}}^{[j]} = \left(\hat{m}^{[j]}(\boldsymbol{x}_t)\right)_{t=1,\ldots,T}$ the vector of function estimates at iteration $j = 1, \ldots, J$, 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 τth sample quantile is a natural choice for quantile regression. The median has also been suggested as a starting value for quantile regression [25].
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 $B = 2p - c$ 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 $\hat{\boldsymbol{m}}^{[j-1]}$:

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

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

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

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 an additive model since each component $\hat{\boldsymbol{u}}^{[j]}$ depends only on one variable k.

C. Base-learners

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 [35]. Since the period-of-day variable has 48 categories, we model it using cyclical P-splines.

P-splines are characterized by a number of parameters: the degree of the B-spline bases, the order of the difference penalty, the number of knots, and the smoothing parameter. Cubic B-splines 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. [36] 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 [31] and [37] 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.

D. Hyperparameters selection

From the different steps described in the gradient boosting algorithm in Section VI-B, 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, [30] 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, J, for which the best value needs to be selected [31].

After setting a range for the number of iterations, (e.g. $J \in \{1, \dots, J_{\text{max}}\}$), the best value of the hyperparameter J can be selected in this range by cross-validation. In particular, we used a holdout approach where each input-output dataset for horizon h is split into training and validation sets. To take into account the dependence between consecutive observations in each dataset, we removed 10 observations before and after each validation point.

VII. Experiments

A. Setup and preprocessing

We focus on day-ahead probabilistic forecasts using the smart meter dataset described in Section II. 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.

The full dataset for residential consumers (without missing values) contains 3639 meters, and for each meter we have access to a time series of length $T = 25728$ comprising half-hourly electricity consumption for almost 18 months. We will consider both disaggregated and aggregated electricity demand; i.e. we will generate probabilistic forecasts for (i) each of the 1000 meters selected from the 3639 meters, and for (ii) 100 other time series obtained by summing the demand of 1000 smart meters randomly selected from the 3639 meters.

We use the first 12 months as a training period to fit the different models, and the remaining data as a test set to evaluate forecast accuracy. The selection of hyperparameters is described at the end of Section VI-D. 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 defined in Section V. In expressions (9) and (10), we use $N \approx 7300$ (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. We make the assumption that Ireland is sufficiently small for the weather at Dublin airport to be similar to the weather elsewhere in the country at any given time. Also, for forecasting we used the actual value of the temperature at Dublin airport. By doing so, note that we do not take into account the uncertainty in the temperature forecasts.

However, since we focus on one-day ahead demand forecasts, the uncertainty in temperature is expected to be small.

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.

To generate probabilistic forecasts at each forecast horizon, we compute the τ-quantiles of the distribution for a set of probabilities $\tau_i = \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$ with $i = 1, \ldots, Q = 21$. To recover the predictive distribution, we use linear interpolation after monotone rearranging to remove quantile crossing. In our gradient boosting procedure, we use a degree of freedom $\text{df} = 4$ 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 [38] available for the R programming language [39]. Our method will be denoted `QR-GAMBOOST` and abbreviated as `QR`.

B. Benchmark methods

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

2) Quantiles computed conditional on time-of-day: This method segments the data by time-of-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 and lagged demands are not accounted for. A variant of this benchmark method has been used in [6] with kernel density estimation methods. This method will be denoted `PeriodOfDay`.

3) Additive models for location and scale parameters of a normal distribution: This method assumes the predictive distributions are normal, and estimates a regression model for both the conditional mean and variance of the distributions. This approach has been considered for probabilistic forecasting of aggregated electricity demand in [40] and [3], with linear and backfitted additive models, respectively. We are considering a variant of this approach, where the regression models are fitted using boosted additive models, i.e. the procedure described in Section VI but with L_2 loss. This method will be denoted `NORMAL-GAMBOOST` and abbreviated as `NORMAL`.

For disaggregated demand, since we are applying a square-root transformation to the data before fitting a normal distribution, this is equivalent to fitting a chi-squared distribution with one degree of freedom to the untransformed data. A square-root transformation has the advantage that it guarantees the non-negativity of the demand forecasts, and can be applied with zeros. Alternative Box-Cox transformations with the normal distribution have been considered for example in wind energy forecasting [17], [11].

C. Results

The first panel of Figure 2 shows the CRPS defined in (10) averaged over all meters for all the forecasting methods over the forecast horizon. The bottom panels show the CRPS, decomposed into the absolute differences and spread components. Figure 3 shows an example of density forecasts for the different forecasting methods.

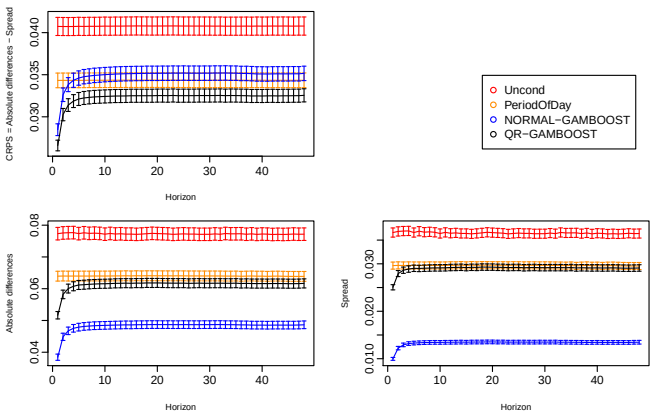

Fig. 2. The CRPS averaged over all meters for the different methods over the forecast horizon decomposed into absolute differences and spread. The error bars give 99.5% confidence intervals for the average CRPS.

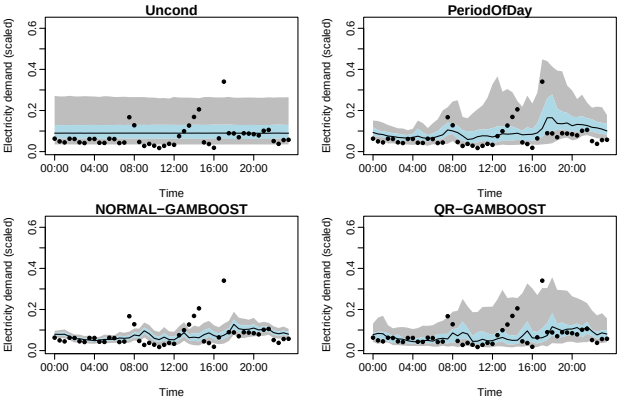

Fig. 3. One-day ahead forecasts for meter 2. The blue and grey regions show 50% and 90% prediction intervals, respectively.

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 that the calendar variables are good predictors of the uncertainty in electricity demand. The remaining panels of Figure 2 show that `PeriodOfDay` has achieved a lower CRPS than `Uncond` by reducing both the absolute differences and spread components. In Figure 3, we see that the predictive distributions 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 variables (used by the two methods) is a good predictor for the first few horizons (see Section III). After the first few horizons, QR has a CRPS close to that of PeriodOfDay, because the recent demand is no longer a good predictor after the first few horizons, but the calendar variables become the main predictors with the highly volatile disaggregated demand.

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 a good approximation for individual electricity consumption.

In the bottom panels of Figure 2, we see that NORMAL has both a lower absolute differences and spread than QR. However, it has a higher CRPS than QR indicating that the predictive distributions of NORMAL are not sufficiently spread out to match the true uncertainty of demand.

By comparing the predictive densities of NORMAL and QR in Figure 3, we see that QR better matches the observations than NORMAL. These results confirm the value of quantile regression methods, allowing greater flexibility for the predictive densities when forecasting individual electricity consumption.

We now compare the methods for aggregated electricity demand, which is less volatile than individual consumption data as shown in Figure 1. Figures 4 and 5 give the same information as Figures 2 and 3, but for 100 aggregated demands, each obtained by summing electricity demand for 1000 smart meters randomly selected from the 3639 meters.

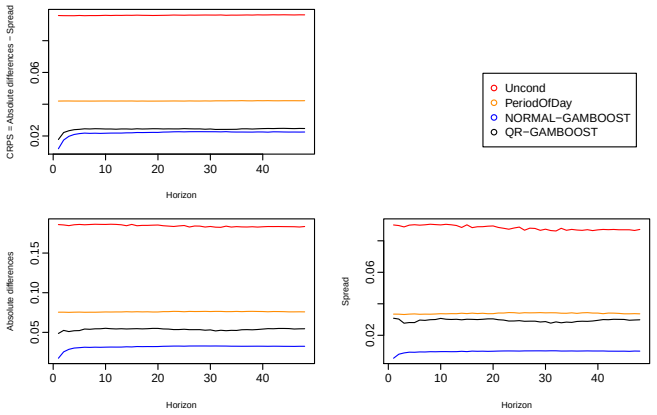

Fig. 4. The CRPS for aggregated demand over the forecast horizon decomposed into absolute differences and spread.

Comparing the CRPS for methods that use lagged demand with those that do not in Figures 2 and 4, we see that the latter methods suffer from a larger drop in performance with aggregated demand compared to disaggregated demand. This is because recent demand is a particularly important predictor for the (smoother) aggregated demand.

Also, in contrast to the results obtained for the disaggregated demand in Figure 2, we see in the top left panel of Figure 4 that NORMAL has a better accuracy showing that the normality assumption is a better approximation for aggregated demand as a consequence of the Central Limit Theorem. This has

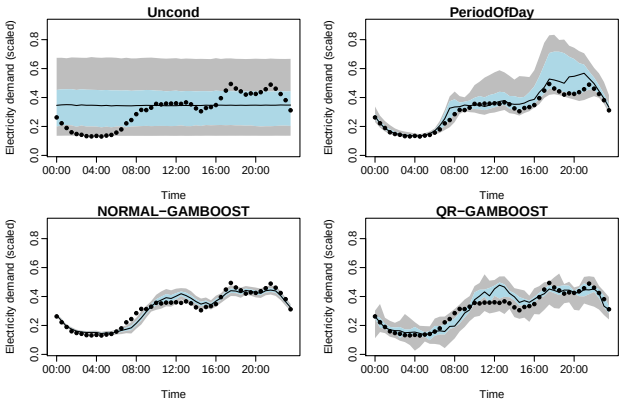

Fig. 5. One-day ahead forecasts for the aggregated demand. The blue and grey regions show 50% and 90% prediction intervals, respectively.

also been observed in [41] for aggregated demand, while [42] and [43] have studied the effect of aggregation on short-term average electricity demand forecasts.

The bottom panel of Figure 4 shows that the predictive densities of NORMAL are relatively sharp compared to other methods, while providing a better CRPS as seen in the top left panel. In Figure 5 we also see that the sharper density forecasts of NORMAL provide relatively good coverage since the demand is much smoother than individual electricity consumption, as illustrated in Figure 3. The better performance of NORMAL with aggregated demand is particularly useful since it has a significantly lower computational load compared to QR.

VIII. 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 proposed to estimate an additive quantile regression model for a set of quantiles of the future distribution using a boosting procedure. By doing so, we can benefit from flexible and interpretable models which also involve an automatic variable selection.

Generating accurate probabilistic time series forecasts is particularly relevant in many energy applications. We have considered the problem of forecasting the distribution of future electricity consumption, on both aggregated and disaggregated scales.

We compared our method with three benchmark methods including a method based on traditional regression, which involves forecasting the conditional mean and variance of the future observations, and making a normality assumption (possibly after a Box-Cox transformation). The results of the comparison between the two methods can be summarized as follows.

At the disaggregated level, with the large variation in consumption behaviour and the high volatility of 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. Also,

because of the high volatility, we found that we can obtain relatively good accuracy simply by conditioning on the period-of-day.

At the aggregated level, where the demand becomes more normally distributed as a consequence of the Central Limit Theorem, normal forecasts have smaller errors than the quantile regression approach. The error decomposition shows that the quantile forecasts lack sharpness; i.e. the predictive distributions are more spread than necessary. Also, because the demand is much smoother, the methods that do not use recent demand suffer from larger errors.

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. Hierarchical probabilistic demand forecasting is another challenging problem for which methods need to be developed. Finally, another important direction is to improve the computational time of probabilistic forecast methods since, in practice, utilities need to deal with millions of smart meters.

REFERENCES

[1] X. Zhu and M. G. Genton, "Short-Term wind speed forecasting for power system operations," *International Statistical Review*, vol. 80, no. 1, pp. 2–23, 2012.

[2] P. Bacher, H. Madsen, and H. A. Nielsen, "Online short-term solar power forecasting," *Solar Energy*, vol. 83, no. 10, pp. 1772–1783, Oct. 2009.

[3] T. K. Wijaya, M. Sinn, and B. Chen, "Forecasting uncertainty in electricity demand," in *AAAI-15 Workshop on Computational Sustainability*, 2015.

[4] T. Hong, "Short term electric load forecasting," Ph.D. dissertation, 2010.

[5] H. Tao and S. Fan, "Probabilistic electric load forecasting: A tutorial review," 2014.

[6] S. Arora and J. W. Taylor, "Forecasting electricity smart meter data using conditional kernel density estimation," *Omega*, 2014.

[7] T. Gneiting and M. Katzfuss, "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, vol. 1, no. 1, pp. 125–151, Jan. 2014.

[8] P. Mirowski, S. Chen, T. K. Ho, and C.-N. Yu, "Demand forecasting in smart grids," *Bell Labs technical journal*, vol. 18, no. 4, pp. 135–158, Mar. 2014.

[9] F. L. Quilumba, W. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE transactions on smart grid*, vol. 6, no. 2, pp. 911–918, Mar. 2015.

[10] Y. Zhang, J. Wang, and X. Wang, "Review on probabilistic forecasting of wind power generation," *Renewable and Sustainable Energy Reviews*, vol. 32, no. 0, pp. 255–270, Apr. 2014.

[11] P. Pinson, "Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions," *Journal of the Royal Statistical Society. Series C, Applied statistics*, vol. 61, no. 4, pp. 555–576, 1 Aug. 2012.

[12] D. M. Bashtannyk and R. J. Hyndman, "Bandwidth selection for kernel conditional density estimation," *Computational Statistics & Data Analysis*, vol. 36, no. 3, pp. 279–298, 28 May 2001.

[13] R. Koenker, *Quantile Regression*, ser. Econometric Society Monographs. Cambridge University Press, 2005.

[14] Commission For Energy Regulation, "Electricity smart metering customer behaviour trials findings report," Dublin: Commission for Energy Regulation, Tech. Rep., 2011.

[15] C. Beckel, L. Sadamori, T. Staake, and S. Santini, "Revealing household characteristics from smart meter data," *Energy*, vol. 78, no. 0, pp. 397–410, 2014.

[16] P. Pompey, A. Bondu, Y. Goude, and M. Sinn, "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, 2014.

[17] J. A. Carta, P. Ramírez, and S. Velázquez, "A review of wind speed probability distributions used in wind energy analysis: Case studies in the canary islands," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 5, pp. 933–955, Jun. 2009.

[18] T. Gneiting, "Quantiles as optimal point forecasts," *International Journal of Forecasting*, vol. 27, no. 2, pp. 197–207, Apr. 2011.

[19] V. Chernozhukov, I. Fernández-Val, and A. Galichon, "Quantile and probability curves without crossing," *Econometrica: Journal of the Econometric Society*, vol. 78, no. 3, pp. 1093–1125, 1 May 2010.

[20] G. Anastasiades and P. McSharry, "Quantile forecasting of wind power using variability indices," *Energies*, vol. 6, no. 2, pp. 662–695, 5 Feb. 2013.

[21] T. Jónsson, P. Pinson, H. Madsen, and H. Nielsen, "Predictive densities for Day-Ahead electricity prices using Time-Adaptive quantile regression," *Energies*, vol. 7, no. 9, pp. 5523–5547, 25 Aug. 2014.

[22] B. Liu, J. Nowotarski, T. Hong, and R. Weron, "Probabilistic load forecasting via quantile regression averaging on sister forecasts," *IEEE transactions on smart grid*, vol. PP, no. 99, pp. 1–1, 2015.

[23] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[24] T. Gneiting, F. Balabdaoui, and A. E. Raftery, "Probabilistic forecasts, calibration and sharpness," *Journal of the Royal Statistical Society. Series B, Statistical methodology*, vol. 69, no. 2, pp. 243–268, 2007.

[25] A. Mayr, T. Hothorn, and N. Fenske, "Prediction intervals for future BMI values of individual children: a non-parametric approach by quantile boosting." *BMC Medical Research Methodology*, vol. 12, p. 6, Jan. 2012.

[26] S. Ben Taieb and R. J. Hyndman, "A gradient boosting approach to the kaggle load forecasting competition," *International Journal of Forecasting*, vol. 30, no. 2, pp. 382–394, Apr. 2014.

[27] S. Fan and R. J. Hyndman, "Short-term load forecasting based on a semi-parametric additive model," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 134–141, 2012.

[28] S. Ben Taieb and R. J. Hyndman, "Boosting multi-step autoregressive forecasts," in *Proceedings of the 31th International Conference on Machine Learning (ICML)*, 2014, pp. 109–117.

[29] T. J. J. Hastie and R. J. J. Tibshirani, *Generalized additive models*. Chapman & Hall/CRC, 1990.

[30] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *Annals of statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[31] P. Bühlmann and B. Yu, "Boosting With the L2 Loss: Regression and Classification," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 324–339, 2003.

[32] R. Koenker, P. Ng, and S. Portnoy, "Quantile smoothing splines," *Biometrika*, vol. 81, no. 4, pp. 673–680, 1 Dec. 1994.

[33] R. Koenker, "Additive models for quantile regression: Model selection and confidence bandaids," *Brazilian Journal of Probability and Statistics*, vol. 25, no. 3, pp. 239–262, Nov. 2011.

[34] R. E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*. The MIT Press, 2012.

[35] T. Kneib, T. Hothorn, and G. Tutz, "Variable selection and model choice in geoadditive regression models," *Biometrics*, vol. 65, no. 003, pp. 626–634, Jun. 2009.

[36] D. Ruppert, "Selecting the number of knots for penalized splines," *Journal of Computational and Graphical Statistics*, vol. 11, pp. 735–757, 2002.

[37] M. Schmid and T. Hothorn, "Boosting Additive Models using component-wise P-Splines," *Computational Statistics & Data Analysis*, vol. 53, no. 002, pp. 298–311, Dec. 2008.

[38] T. Hothorn, P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner, "Model-based boosting 2.0," *Journal of Machine Learning Research: JMLR*, vol. 11, pp. 2109–2113, 2010.

[39] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015.

[40] R. Engle, C. Granger, R. Ramanathan, and F. Vahid-Araghi, "Probabilistic methods in forecasting hourly loads," Electric Power Research Inst., Palo Alto, CA (United States); Quantitative Economic Research, Inc., San Diego, CA (United States), Tech. Rep. EPRI-TR-101902, 1 Apr. 1993.

[41] R. Sevlian, S. Patel, and R. Rajagopal, "Distribution system load and forecast model," 11 Jul. 2014.

[42] Y.-H. Hsiao, "Household electricity demand forecast based on context information and user daily schedule analysis from meter data," *Industrial Informatics, IEEE Transactions on*, vol. 11, no. 1, pp. 33–43, Feb. 2015.

[43] R. Sevlian and R. Rajagopal, "Scaling law of very short term electricity load forecasting on varying levels of aggregation," 2014.