ARTICLE IN PRESS
International Journal of Forecasting ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

International Journal of Forecasting

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

GEFCom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation

Jingrui Xie *, Tao Hong

University of North Carolina at Charlotte, Charlotte, NC, USA

ARTICLE INFO

Keywords:
Probabilistic forecasting
Load forecasting
Residual simulation
Regression analysis
Time series modeling
Neural networks
Forecast combination

ABSTRACT

We present an integrated solution for probabilistic load forecasting. The proposed solution was the basis for Jingrui Xie's submission to the probabilistic load forecasting track of the Global Energy Forecasting Competition 2014 (GEFCom2014), and consists of three components: pre-processing, forecasting, and post-processing. The pre-processing component includes data cleansing and temperature station selection. The forecasting component involves the development of point forecasting models, forecast combination, and temperature scenario based probabilistic forecasting. The post-processing component embodies residual simulation for probabilistic forecasting. In addition, we also discuss several other variations that were implemented during the competition.

1. Introduction

Utilities have been using point forecasts for system and financial planning for several decades. However, the worldwide modernization of the power grid has meant that the electricity demand has become more versatile and less predictable than ever before. As a result, probabilistic load forecasts have become increasingly important in helping utilities to quantify the uncertainties in the electricity demand. A recent tutorial review on probabilistic load forecasting (Hong & Fan, this issue) summarized the significant developments in this field and introduced a range of techniques for producing and evaluating probabilistic load forecasts.

In an attempt to address and overcome the challenges of probabilistic energy forecasting, the IEEE Working Group on Energy Forecasting organized the Global Forecasting Competition 2014 (GEFCom2014) (Hong et al., this issue). The probabilistic load forecasting track of GEFCom2014 provided the competitors with six years (2005–2010) of hourly load data and 10 years (2001–2010) of hourly temperature data from 25 anonymous weather stations. During each of the 15 weeks of the competition, each participating team was asked to provide a one-month-ahead probabilistic load forecast in the form of quantiles. The actual load and temperature data from the previously forecasted month was released incrementally every week. The detailed competition rules, a description of the data and a summary of the methods of selected teams are provided by Hong et al. (this issue). This paper presents Jingrui Xie's methodology in detail.

The methods and models implemented by Jingrui Xie evolved over the course of the competition. In this paper, we focus on the core solution framework, which was the basis of Jingrui Xie's implementations during GEFCom2014. As Fig. 1 shows, the solution framework consists of three components:

- The *pre-processing* component, which includes two parts. The first part uses Tao's vanilla benchmark model

* Corresponding author.
E-mail addresses: jingrui.rain.xie@gmail.com (J. Xie), hongtao01@gmail.com (T. Hong).

http://dx.doi.org/10.1016/j.ijforecast.2015.11.005

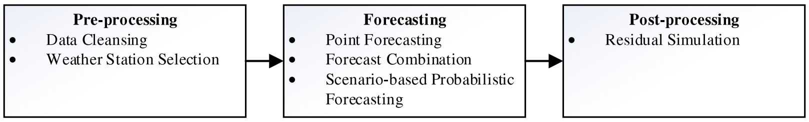

Fig. 1. Framework of Jingrui Xie's solution in GEFCom2014.

(Hong, Pinson, & Fan, 2014; Hong, Wang, & Willis, 2011) for outlier detection and data cleansing. The second part uses the same model for weather station selection, following the methodology proposed by Hong, Wang, and White (2015).

- The *forecasting* component consists of three steps. The first step sharpens the underlying point forecasting models; the second step combines the point forecasts; and the third step generates temperature scenario based probabilistic load forecasts based on the point forecasting model and 10 years of temperature history.
- The *post-processing* component simulates the residuals of the selected point forecasting models from the forecasting component, in order to improve the probabilistic forecast further.

The next three sections of this paper will discuss these three components. In Section 5, we will discuss alternative implementations and the results. The paper will then be concluded in Section 6.

2. Pre-processing

2.1. Data cleansing

The data cleansing process starts with the benchmark model in Eq. (1). It is a multiple linear regression (MLR) model with the following main and cross effects:

i. main effects: a chronological trend variable (*Trend*), the 1st to 3rd order polynomials of the temperature (T_t, T_t^2 and T_t^3), and the class variables *Month*, *Weekday* and *Hour*;

ii. cross effects: $Hour_t * Weekday_t$, $T_t * Month_t$, $T_t^2 * Month_t$, $T_t^3 * Month_t$, $T_t * Hour_t$, $T_t^2 * Hour_t$, and $T_t^3 * Hour_t$.

After estimating the benchmark model using all of the historical data, we calculate the absolute percentage error (APE) for each hourly load observation. The observations with APE values of greater than 50% are treated as outliers. For these outliers, we then replace the original observations with the predicted values from the benchmark model. This data cleansing process modifies about 0.05% of the historical load data. As an example period, Fig. 2 shows October 2–3, 2006, which this method detects as a series of outliers.

$$\begin{aligned} E(Load_t) = {} & \beta_0 + \beta_1 * Trend_t + \beta_2 * T_t + \beta_3 * T_t^2 \\ & + \beta_4 * T_t^3 + \beta_5 * Month_t + \beta_6 * Weekday_t \\ & + \beta_7 * Hour_t + \beta_8 * Hour_t * Weekday_t \\ & + \beta_9 * T_t * Month_t + \beta_{10} * T_t^2 * Month_t \\ & + \beta_{11} * T_t^3 * Month_t + \beta_{12} * T_t * Hour_t \\ & + \beta_{13} * T_t^2 * Hour_t + \beta_{14} * T_t^3 * Hour_t. \qquad (1) \end{aligned}$$

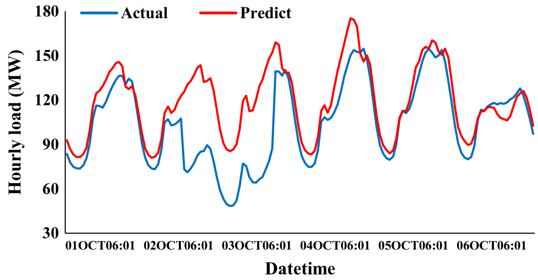

Fig. 2. Hourly actual and predicted loads for Oct. 1–6, 2006.

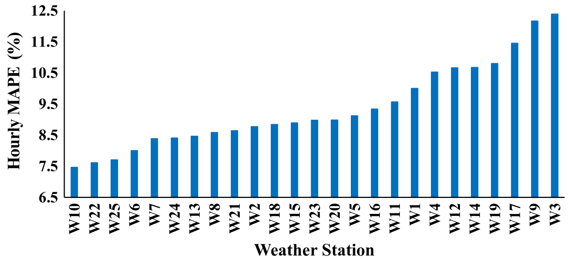

Fig. 3. Weather stations ranked by ascending MAPE (%).

2.2. Weather station selection

Hourly temperatures from 25 anonymous weather stations (W_1–W_{25}) were provided in the competition, but the competition organizer did not release any geographic information for the individual weather stations. Thus, we select the appropriate weather stations for the utility by following the weather station selection method proposed by Hong et al. (2015).

We begin by feeding the benchmark model with the temperature series from the 25 weather stations. Estimating the model using the data from the years 2007 to 2009, we obtain 25 sets of in-sample fit results, one from each of the 25 weather stations. We then calculate the mean absolute percentage errors (MAPE) of the in-sample fit results, and sort the weather stations by MAPE values in ascending order, as Fig. 3 shows.

By averaging the hourly temperature series of the top n weather stations, we can obtain combined weather stations, denoted by CW_n. For example, CW_3 is the hourly temperature series created by averaging the hourly temperatures of W_{10}, W_{22} and W_{25}. We then feed the benchmark model with each of the 25 combined temperature series in turn. Using the data from the years 2007–2009 to estimate the model, we can then forecast the year 2010. Fig. 4 shows the MAPE values of the combined weather stations for the year 2010. The best accuracy occurs for CW_{11} (i.e., the average temperature of the top 11 weather stations), which yields the lowest MAPE value of 6.18%.