A New Approximation Method for Generating Day-Ahead Load Scenarios

Yonghan Feng, Dinakar Gade, and Sarah M. Ryan
Iowa State University
Ames, IA, USA
{yhfeng,dgade,smryan}@iastate.edu

Jean-Paul Watson
Sandia National Laboratories
Albuquerque, NM, USA
jwatson@sandia.gov

Roger J-B Wets and David L. Woodruff
University of California Davis
Davis, CA, USA
{rjbwets,dlwoodruff}@ucdavis.edu

Abstract— **Unit commitment decisions made in the day-ahead market and resource adequacy assessment processes are based on forecasts of load, which depends strongly on weather. Two major sources of uncertainty in the load forecast are the errors in the day-ahead weather forecast and the variability in temporal patterns of electricity demand that is not explained by weather. We develop a stochastic model for hourly load on a given day, within a segment of similar days, based on a weather forecast available on the previous day. Identification of similar days in the past is based on weather forecasts and temporal load patterns. Trends and error distributions for the load forecasts are approximated by optimizing within a new class of functions specified by a finite number of parameters. Preliminary numerical results are presented based on data corresponding to a U.S. independent system operator.**

Index Terms— **Demand forecasting, Load modeling, Power system planning, Stochastic processes.**

I. Introduction

Constraints on the operation of thermal generating units require them to be committed well in advance of when they may be needed to provide power. Typically, scheduling decisions for a day *D* are made on day *D*-1 according to forecasts of hourly load aggregated across the buses in a load zone. The information available to planners on day *D*-1 includes weather forecasts for day *D* and historical records of previous weather forecasts combined with the corresponding actual hourly loads. The historical data show temporal variation in the load over a day that varies according to season of the year and day of the week. While some patterns in temporal load are predictable based on business hours and diurnal light patterns, the portion of load derived from heating and cooling depends strongly on the weather. Further, although numerical weather prediction models have become increasingly accurate, there remains uncertainty associated with the day-ahead weather forecasts. Thus, the challenge for planners is to form an accurate picture of the day-ahead load, which not only includes point forecasts of the load in each hour, but also acknowledges the precision, or lack hereof, associated with those forecasts.

The uncertainty associated with day-ahead scheduling is gaining increased attention with the growing penetration of variable generation resources and demand response mechanisms. These developments augment the uncertainty that has always existed in both the load and the availability of thermal units. These factors have motivated investigations into stochastic optimization methods for unit commitment, which require probabilistic descriptions of the supply and demand for power several hours in advance of when they will be realized. This paper describes a novel method to develop a stochastic model for the load on day *D* based on a weather forecast available on day *D*-1. The resulting probabilistic description of demand is designed to be combined with corresponding stochastic models for variable generation and resource availability that, together, will comprise inputs for a stochastic unit commitment optimization. To facilitate implementation, our method is similar to a load forecasting approach already used by system planners and is compatible with a wind power forecasting method developed by some commercial firms.

Section II very briefly describes major load modeling approaches and Sections III and IV describe our approaches for identifying similar days and approximating load patterns. We present some preliminary numerical results in Section V and conclusions in Section VI.

II. Load Forecasting Approaches

Common methods for short-term (hour- to week-ahead) load forecasting can be characterized as either artificial intelligence or statistical techniques [1]. Artificial intelligence methods, such as artificial neural networks, are widely used but do not provide probabilistic information that could be used

This material is based on work sponsored by the U.S. Department of Energy under the ARPA-E Green Energy Network Integration (GENI) Program. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

to generate multiple probability-weighted scenarios. Among statistical approaches, the most prevalent methods are time series and regression models. Due to limited space, we do not provide a complete review but refer the reader to recent surveys such as [2]. Instead, we highlight recent representative samples of statistical approaches.

Exploratory data analyses indicate that electricity load in a region depends on time of year, day of the week, and hour of the day; and is strongly influenced by weather. Forecasting methods vary in how they account for these factors. They are typically evaluated according to the accuracy of the point forecasts they provide for each hour of the next day.

The weather variable most commonly used to predict load is temperature because, in most parts of the industrialized world, the peak load occurs in summer due to air conditioning. Temperature also affects load in winter, but in the opposite direction due to electrical heating. Humidity increases load in the summer, while cloud cover increases load in the winter and reduces it in the summer. However, these effects are much smaller than that of temperature. Liu et al. [3] noted the nonlinear relationship between temperature and load when all hourly data throughout a year were considered together and applied a nonparametric regression method to estimate it. They fit a time series model to the residuals of the load-temperature regression and included lags of 1, 24, and 168 hours in their day-ahead forecasting model. Using actual historical temperatures and loads obtained from a U.S. utility, they achieved a post-sample mean absolute percent error (MAPE) of 1.2% for their 24-hour-ahead forecasts. However, they did not explain how they estimated the 23-hour-ahead load values required by the lag 1 term. A drawback of time series methods, when used to forecast more than one step ahead, is that uncertainty propagates through the lagged terms, distorting the variability of forecasts for remote time periods. This renders them less suitable for building a stochastic process for the load.

Hong et al. [1] developed a multiple linear regression model that included, as independent variables, a piecewise-quadratic function of temperature; dummy variables to represent hour, day-type, and month; a linear trend; and interactions among these variables. They obtained a post-sample MAPE of 4.6% when using actual weather data to hindcast hourly loads for a U.S. utility over a one-year period.

Black [4] also used multiple linear regression to examine the influence of weather on load but focused on summer weekdays in the region served by ISO-New England. He incorporated time-of-day effects by developing a separate regression model for each hour of the day, including as independent variables temperature, humidity, solar radiation; lagged, averaged, squared and cubed values of these variables; and interactions among them. The out-of-sample MAPEs yielded by these models averaged 2-3% for the whole New England region and 3-4% for individual subregions such as Connecticut and Southeast Massachusetts.

While hindcasting studies that use historical weather data as input are useful for identifying factors and relationships that affect hourly loads, they do not assess the accuracy or precision of the load forecasts available in practice, which necessarily rely on day-ahead weather forecasts. Although weather prediction has greatly improved in recent years, day-ahead forecasts remain imperfect. An alternative approach to short-term load forecasting is to identify similar days within a historical database, where the similarity is based on weather, day of the week and time of year. For example, ISO-New England identifies up to five similar days drawn from the same season with the same day-type according to similarity of their actual temperatures to the forecast temperature of the given day as well as similarity of forecast loads in the last hour of the previous day [5]. Our method has some commonality with this approach, in that we create segments of days that are similar in some sense. Then, within each segment we employ a functional regression method to approximate the probability distribution of load in each hour of the day ahead.

III. Data Segmentation

We use a multi-step procedure to control for time of year and type of day, and then approximate the relationships between weather forecast and distribution of hourly load sequences within segments of similar days. Starting from a historical database of day-ahead hourly weather forecasts and corresponding actual load sequences, the steps are as follows:

1. Identify date ranges, or “seasons,” in which the relationship between weather and load, disregarding day-of-week effects, is likely to be similar. This is an ad hoc characterization that should vary by region and account for diurnal light patterns, heating vs. air conditioning, and sociological factors such as holiday lighting and schools being in session or not.

2. Within each season,

a. Compute the average load for each hour in each day of the week, considering holidays as Sundays. Then compute multipliers for each day of the week to transform that day’s load sequence to a Wednesday load sequence.

b. Using the transformed load sequences, identify segments of days in which the relationship between the day-ahead weather forecast and the actual load was similar. Within each segment, approximate this relationship as a regression function and also approximate the distribution of residuals from the regression. The approximation method is described in Section IV.

Having completed the segmentation and approximation steps, the procedure for generating a distribution of load sequences for a given day, *D*, is:

1. Identify the season to which day *D* belongs and the segment to which its weather forecast, generated on day *D*-1, belongs.

2. Apply the regression function to the weather forecast and then, if necessary, transform the expected (Wednesday) load sequence to match the day of the week.

3. Add a randomly distributed error according to the estimated error density for the segment.

IV. Approximation Method

Within each segment, we construct a stochastic process for the next day's load in two steps. We begin by deriving an estimate of the next day's (*D*) load pattern by means of a functional regression based on the hourly sequences of temperature and dewpoint forecasts on day *D*-1. Next, we use the errors that result from applying this model, within the segment of days that were used to train the regression function, to estimate an error density function for each hour. Combining these two estimates allows us to build a stochastic process from which we can generate potential load scenarios for day *D*.

A. Fit Approximating Function

To obtain the regression curve we rely, for the first time, on a new technology based on representing functions, more precisely approximating them as epi-splines. An *epi-spline* of order *k* is a real-valued function where the interval on which it is defined has been partitioned into a large but finite collection of subintervals and requiring that on each subinterval, its *k*th derivative be constant. Epi-splines were originally used to derive term and volatility structures associated with financial markets [6] and since have been used in a number of other contexts. We exploit the fact that epi-splines are defined by a finite number of parameters, namely, the *k*th derivatives and $k-1$ integration constants, and are particularly suited to converting an infinite dimensional functional estimation problem, equivalently an optimization problem, into a finite dimensional one. In addition, epi-splines allow the incorporation of constraints that express "soft" information about the shape of the regression function [7]. For example, by requiring that the expressions defining the first derivative be nonpositive, we can impose the condition that during certain time spans, the load necessarily decreases, and so forth.

Dividing the time interval (0, 24] hours into $N = 24m$ subintervals, our regression function takes the form:

$$l_h^j = \sum_w s_w(mh) w_h^j + e_h^j\,, \tag{1}$$

where j denotes a day in segment *J*, $h \in \{1, \dots, 24\}$ denotes an hour of the day, *w* indexes a set of weather variables, w_h^j denotes a forecast of weather variable *w* for hour *h* of day *j*, and e_h^j denotes the regression error. We rely on 2nd order epi-splines $s_w(\cdot)$ generated on both temperatures and dewpoints during the summer months. Let $\delta = 1/m$. In general, for the *k*th subinterval $(t_{k-1}, t_k]$ and τ belonging to this subinterval, a second-order epi-spline is defined as:

$$s(\tau) \equiv s_0 + v_0\tau + \delta \sum_{i=1}^{k-1} (\tau - t_i + \delta/2) a_i$$
$$+ (1/2)(\tau - t_{k-1})^2 a_k\,. \tag{2}$$

The function $s(\tau)$ is completely determined once we fix the integration constants s_0, v_0 and the 2nd derivatives $(a_1, \dots, a_N)$. Finding the regression curve then comes down to finding this finite number of parameters to minimize some measure of the errors in estimating the loads observed in a particular segment, taking into account any side constraints (soft information) that one may have included in the formulation of the estimation problem. Applying the L_1 norm (minimizing the sum of absolute errors) or the L_∞ norm (minimizing the maximum error) in the objective results in a linear program, while the L_2 norm (sum of squared errors) leads to a quadratic program with linear constraints.

The quality of the approximation is higher if segments of historical days are homogeneous, but could suffer if a tight segmentation results in a very small training data set for each segment. Decision makers are often very concerned that peak loads not be underestimated, and such constraints can be incorporated as soft information using linear inequalities. If there is a high degree of confidence about the forecast for load just before midnight on day *D*-1, then the approximated load just after midnight on day *D* can be constrained to fall within a certain range. Experimentation with incorporating such considerations is ongoing, but the results reported in Section V are derived without any soft information constraints.

B. Approximate Error Distributions

Once the regression curve has been determined, which can be interpreted as providing the "overall trend" or expected value of the stochastic load process (for day *D*), our next step is to generate distributions of the errors, again for a given day-type segment. For each hour, we generate a nonparametric estimate of this distribution by approximating the density by an exponential epi-spline; i.e., a function of the type

$$f(x) = e^{-u(x)},\ x \in [\alpha, \beta] \tag{3}$$

where *u*(*x*) is a 2nd order epi-spline of the form defined in Section IV.A and the range [α, β] represents a number (say 3) of sample standard deviations around the sample mean of errors computed for the segment. We can include any soft information we might have or suspect about the shape of this distribution, for example, we can stipulate that the density be unimodal. We use maximum likelihood as our criterion function and the estimation problem reduces to solving a finite dimensional optimization problem with linear constraints and a slightly nonlinear but convex objective. Extensive experimentation regarding density estimation is reported in [8] and further theoretical foundations are laid out in [9].

V. Preliminary Results

We obtained historical hourly loads for the eight load zones in ISO-New England from January, 2006, through August, 2012, as well as the corresponding day-ahead hourly forecasts of temperature and dewpoint temperature from March, 2007, through August, 2012. Because, in the aftermath of the financial crisis, the load was about 5% lower in 2009-11 than in 2006-08, we conduct our analysis using the more recent data. In this paper, we present results for the Connecticut load zone, which accounts for about 26% of the total demand in New England. We focus on summer days, defined as June – September, in which the highest peak loads of the year occur in the late afternoon. Figure 1 shows the average hourly load sequence for each summer day of the week.

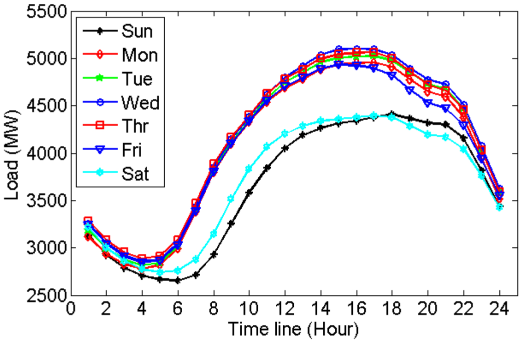

Figure 1. Average load sequence in each day of summer, 2010-2012

The process for identifying segments of similar days is a subject of ongoing study. Figure 2 shows scatter plots for hourly load vs. forecast temperature in the summer months of 2010 – 2012. Noting the strong positive correlation, we used k-means with Euclidean distances to cluster the sequences of hourly forecast temperatures into k=3 sets. Figure 3 plots the centroid of each set as a 24-hour time series. The centroid for all summer days nearly coincides with the moderate day centroid. This process, applied to data from 2010 to 2012, resulted in identifying 3 segments of 91 hot days, 152 moderate days, and 84 cool days, respectively. Figure 4 shows the relationship between load and dewpoint temperature, in addition to (dry bulb) temperature.

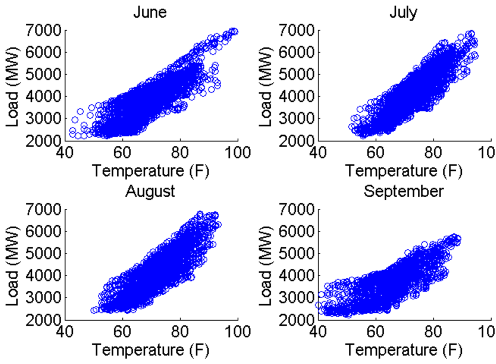

Figure 2. Load vs. day-ahead temperature forecast in the four summer months, 2010-12.

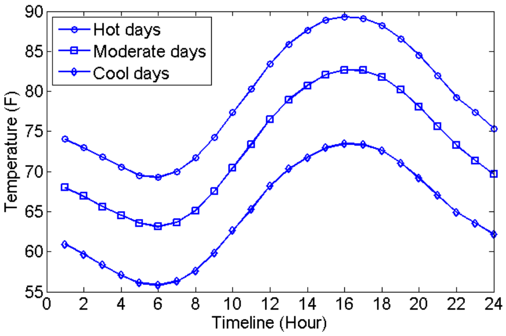

Figure 3. Centroids of the clusters of summer day temperature forecasts.

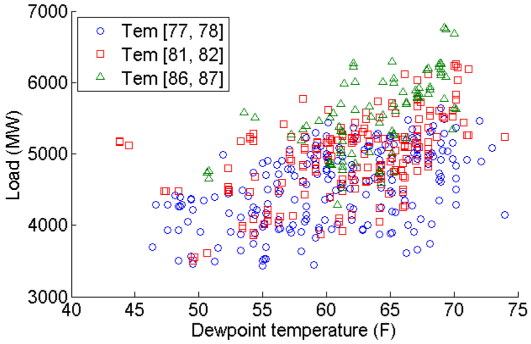

Figure 4. Load vs. forecast dewpoint temperature for hours with forecast temperatures in specified ranges.

We developed the approximating functions using data from 2010 and 2012 and applied them to hindcast loads in 2011, using $m = 1$ subinterval for each hour. All results presented here are based on the L_1 norm. Note that, unlike previous hindcasting studies, the estimated loads in 2011 were obtained by feeding the historical day-ahead weather *forecasts* into the models; thus, some of the hindcast error is attributable to weather forecast error and some to our model. Figure 5 shows within-sample relative errors in the load for each of the three segments. For each summer day in 2011 (the test set), we determined which segment it belonged to by minimizing the distance between its forecast temperature sequence and each of the three centroids, then applied that segment's model. Figure 6 shows the corresponding post-sample relative errors for the three segments in 2011. The within-sample and post-sample MAPEs are given in Table I.

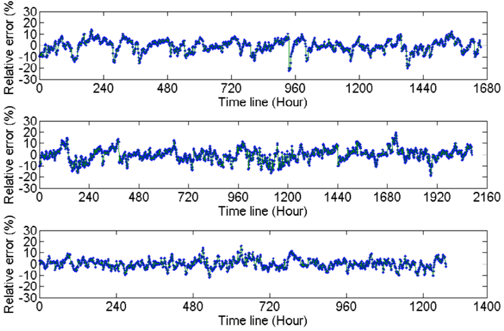

Figure 5. Time series plots of errors in hot, moderate, and cool summer days (from top to bottom) of the training set.

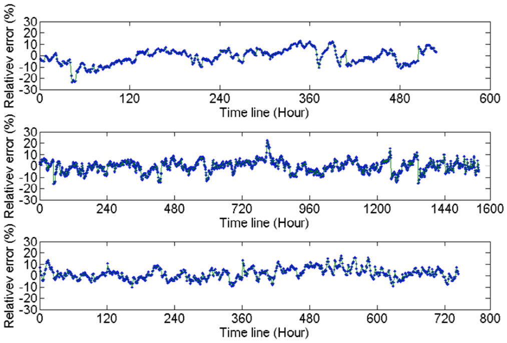

Figure 6. Time series plots of errors in hot, moderate, and cool summer days (from top to bottom) of the test set

TABLE I. MAPE (%) for each segment.

Segment	Training set	Test set
Cool days	4.11	5.40
Moderate days	4.13	3.98
Hot days	3.03	4.01

For generating alternative scenarios of load sequences, the error distributions provide a measure of variability. Figure 7 shows approximate error densities for hours 6 and 17 of hot summer days, when the lowest and highest loads occur. Peak loads (hour 17) are more uncertain. These densities can be numerically integrated to obtain cumulative distribution functions, from which multiple scenarios of load for each hour, corresponding to a given weather forecast, can be generated.

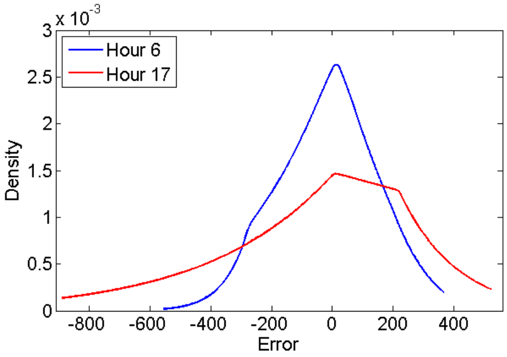

Figure 7. Approximate densities for the errors in the trough and peak hours.

VI. CONCLUSIONS

We have described new methods for obtaining distributions of load; i.e., predictions of load uncertainty, in each hour of day *D*, based on weather forecasts available on day *D* - 1. Our goal is to use the estimated trends and error distributions to generate probabilistic scenarios for the day-ahead load to use in stochastic unit commitment procedures. Another major and growing source of uncertainty in unit commitment is introduced by renewable generation, such as wind power. Recently, some commercial forecasts for wind power are being generated based on identifying analog weather patterns. By estimating our models within segments of similar days, we hope to facilitate the development of joint distributions of load and wind power based on weather forecasts.

Our experiments show that the models produce errors that are competitive in the aggregate. In fact, we obtain similar MAPE values as have been found in hindcasting studies that eliminated weather forecast uncertainty. We are continuing to refine the methods. On-going research focuses on reducing the errors in the most important peak load periods and adapting the approach to times of year when temperature is not as strong a predictor of the load.

ACKNOWLEDGMENT

The authors are grateful to Bill Callan, Joe Mercer and Eugene Litvinov of ISO-New England for helpful discussions.

REFERENCES

[1] T. Hong, M. Gui, M. E. Baran, and H. L. Willis, "Modeling and forecasting hourly electric load by multiple linear regression with interactions," in *Power and Energy Society General Meeting*, Minneapolis, MN, 2010.

[2] E. A. Feinberg and D. Genethliou, "Load forecasting," in *Applied Mathematics for Restructured Electric Power Systems*, J. H. Chow, F. F. Wu, and J. Momoh, Eds., ed: Springer, 2005, pp. 269-285.

[3] J. M. Liu, R. Chen, L.-M. Liu, and J. L. Harris, "A semi-parametric time series approach in modeling hourly electricity loads," *Journal of Forecasting*, vol. 25, pp. 537-559, 2006.

[4] J. D. Black, "Load Hindcasting: A Retrospective Regional Load Prediction Method using Reanalysis Weather Data," M.S., Mechanical and Industrial Engineering, University of Massachusetts Amherst, 2011.

[5] ISO-New England, "Create demand forecast," http://www.iso-ne.com/rules_proceds/operating/sysop/out_sched/sop_outsch_0040_0010.pdf, 2011.

[6] R. Wets and S. Bianchi, "Term and volatility structures," in *Handbook of Asset and Liability Management*, S. Zenios and W. Ziemba, Eds., ed: Elsevier, 2006, pp. 26-68.

[7] G. Pflug and R. Wets, "Shape restricted nonparametric regression with overall noisy measurements," *Journal of Nonparametric Statistics*, forthcoming, 2012.

[8] R. Wets. (2009). *Fusion of hard and soft information: density estimation*. Available: http://www.math.ucdavis.edu/~prop01/

[9] J. Royset and R. Wets, "Nonparametric density estimation with soft information using exponential epi-splines," University of California Davis, 2012.