Energy 133 (2017) 108–114

Contents lists available at ScienceDirect

Energy

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

Short-term load forecasting using a two-stage sarimax model

Agostino Tarsitano*, Ilaria L. Amerise

Dipartimento di Economia, Statistica e Finanza, Università della Calabria, Via Pietro Bucci, Cubo 1c, 87036 Rende, CS, Italy

ARTICLE INFO

Article history:
Received 12 October 2016
Received in revised form
5 May 2017
Accepted 21 May 2017
Available online 23 May 2017

Keywords:
Model building
Time series
Linear regression
External predictors

ABSTRACT

The primary aim of this study is to develop a new forecasting system for hourly electricity load in six Italian macro-regions. The statistical methodology is centered around a dynamic regression model in which important external predictors are included in a seasonal autoregressive integrate moving average process (sarimax). Specifically, the external variables are lagged hourly loads and calendar effects.

We first use backward stepwise regression to estimate regression parameters and obtain residual series. We then identify an optimal sarima process for the residuals and, finally, the parameters of the regression and of the time series models are jointly estimated using a sarimax process selected within a small set of variants of the sarima model found for the residuals.

One-day and nine-day ahead prediction performance of the proposed methodology show that intelligent integration of linear regression, time series and computational resources into a unique approach may provide accurate predictions for short-term electric loads.

1. Introduction

Accurate short-term load forecasting (STLF) has great importance in power system planning operation and control where it is used to align the energy demand to energy generation and energy purchase in the competitive energy markets. Large variation in forecast accuracy might cause serious financial penalties and/or missing valuable opportunities for profit. Consequently, any reduction in forecast errors will result in a boost to the system so as to operate closer to the optimum capacity.

Bunn [5] notes that electricity markets are structurally different to other commodities, and the real-time dynamic balancing of the electricity network involves many external factors. Because of this, it is not a simple matter to transfer conventional models of time series analysis to wholesale power markets.

Over the years, several methods have been developed to model the electricity load. See Weron [36] and Hong [15] for a review. There has been a recent increase in literature, also under the impulse of Hong et al. [16]. In general terms, the various methods of load forecasting that can be applied at the macro level give rise to two main classes of procedures: first, statistical techniques such as seasonal autoregressive moving average, linear regression, non-parametric e semi-parametric regression, periodic stochastic time series, general exponential smoothing, state space model and Kalman filter analysis (*e.g.*, [24,25,37]; second, artificial intelligence based techniques such as support vector machines fuzzy logic, artificial neural networks and expert based systems (*e.g.* Refs. [20,32–34]. In the present paper, we describe and test a methodology based on linear statistical methods for STLF. An overview on this theme is provided in Feinberg & Genethliou [12] and Almeshaiei & Soltan [1].

Various techniques have been investigated to solve the problem of short term forecasting, with varying degrees of success and there are widespread studies making efforts for improving the accuracy of prediction. Considerable attention has been directed toward the multiple linear regression model.

$$L_t = \beta_0 + \sum_{j=1}^{m} \beta_j X_{t,j} + e_t \ . \tag{1}$$

where L_t is the hourly energy demand expressed in *MWh*; $X_{t,j}, j = 1, 2, \cdots, m$ are independent variables at time t that represent the most relevant variables influencing the demand of electricity together with the dummy variables that capture various features and special conditions concerning consumer habits. The parameter β_j measures how the load L_t is related to the j-th variable $X_{t,j}$. The intercept term β_0 can be interpreted as the expected load (in *MWh*) when all the non-dummy variables in (1) have been rescaled to

* Corresponding author.
E-mail addresses: agostino.tarsitano@unical.it (A. Tarsitano), ilaria.amerise@unical.it (I.L. Amerise).

http://dx.doi.org/10.1016/j.energy.2017.05.126
0360-5442/

have zero mean and the dummy variables have a value of zero. Finally, e_t is a random error with zero mean and constant variance.

Often it is assumed that the errors $e_1, e_2, \cdots, e_n$ form a white noise process, although it is quite plausible that the residuals are strongly correlated. When disturbances exhibit serial correlation, least squares yield unbiased, but inefficient estimators, thus invalidating all tests of significance. The remedy is to utilize a model that takes into account the correlation structure in the residuals.

In the past few years, a class of time series models named sarimax (seasonal autoregressive integrated moving average with external or exogenous regressors) appear to be very popular in short-term load forecasting. See, for example, Engle et al. [11]; Rothe et al. [30]; Chikobvu & Sigauke [7].

The general form of a sarimax model can be written as

$$L_t = \beta_0 + \sum_{j=1}^{m} \beta_j X_{t,j} + [\phi^*(B)]^{-1}\theta^*(B)a_t \ . \tag{2}$$

where B is the usual backward shift operator $B^j z_t = z_{t-j}$, the integer s is the seasonal period ($s = 24$ in our case) and a_t are independent and identically distributed random residuals with zero mean, variance σ_a^2 and finite kurtosis. The error e_t appears implicitly as $e_t = [\phi^*(B)]^{-1}\theta^*(B)a_t$. In practice, the hypothesis of a white noise process is placed on a_t rather that on e_t. Moreover

$$\begin{aligned}\phi^*(B) &= 1 - \phi_1^* B - \phi_2^* B^2 - \cdots - \phi_{p^*}^* B^{p^*}; \quad \theta^*(B) \\ &= 1 - \theta_1^* B - \theta_2^* B^2 - \cdots - \theta_{q^*}^* B^{q^*} \ .\end{aligned} \tag{3}$$

are polynomials in B. Some of the parameters may be zero or otherwise constrained, so that (3) could be a multiplicative seasonal model. The other characteristics of $\phi^*(B)$ and $\theta^*(B)$ will be discussed in section 3.1.

In the present paper, we study a linear regression model in which lagged values of the dependent variable occur as regressors and the disturbances follow a sarima process (see Refs. [26–28]. What distinguishes our approach from the conventional sarimax method is, firstly, that, in our case, the sarimax model is preceded by the filtering of regression errors and by an identification stage devoted to determine an optimal sarima process that represents them effectively. Secondly, that search of the optimal orders is carried out by a brute-force algorithm running over a very varied set of feasible processes. Finally, the length of the time series dealt with is of the order of tens of thousands (possibly, hundreds of thousands) of value. This puts a heavy strain on computational resources and algorithms and renders acute the need for a parsimonious and effective model to anticipate the hourly electricity load. The combination of the three-model building phases constitutes the specific contribution of this paper.

The structure of the paper is as follows. Section 2 proposes a multiple linear regression model to forecast hourly electricity consumption by using a specific selection of variables that influence the load in each of the six Italian macro-regions considered in this paper. The regression errors are assumed to be white noise processes. Section 3 attempts to overcome the limitations associated with the assumption of white noise. To this end, three types of dynamic regression model are analyzed: sarimax, reg-plus-sarma and two-stage sarimax. In Section 4 a numerical study is performed to show the gradual improvement of forecasting accuracy as the model changes from the first to the last. Conclusions and indications for future research are given in Section 5.

2. STLF based on linear regression

Linear regression within the framework of STLF is used to model, as closely as possible, the relationship between energy consumption and other relevant regressors. After a series of preliminary experiments and literature reviews on the regressors, we have restricted the set of explanatory variables to lagged effects (because the load at a given hour is dependent above all on previous loads) and calendar effect variables, that is, dummy variables that capture features and special conditions concerning consumer habits. Applications of regression models for Italian electricity data are presented, for example, in Gori & Takanen [14] and Bianco et al. [4].

The resulting equation for the multiple linear regression is

$$L_t = \beta_0 + \sum_{i=1}^{m} \beta_i D_{t,i} + \sum_{j=1}^{k} \beta_{m+j} L_{t-j} + e_t \ . \tag{4}$$

where e_t are random errors with mean zero and constant variance, k is the furthest lag with a nonzero coefficient and $D_{t,i}, i = 1, 2, \cdots, m$ is a set of dummy variables. We have also explored, although with limited success, the influence of predictors linked to hourly temperatures. Evidently, their characteristics are already incorporated into the dummies and/or into the lagged variables. On this, see Soares & Medeiros [31].

As we mentioned earlier, our goal in this paper is to set up a scheme that will generate a solution to (4). The first step of our approach is the formulation of a regression equation with errors described by a rudimentary white noise $(0,0,0) \times (0,0,0)_{24}$ process, which allows us to estimate the regression parameters by ordinary least squares. It follows that the only problem to be solved is to choose the regressors. In this regard, we use the backward stepwise selection procedure to identify lags and dummies felt to be statistically important. The technique starts with a regression equation consisting of all the explanatory variables that are potentially associated with the hourly energy demand. The variables are sequentially eliminated from the model one at a time until no more predictors can be removed. At any stage of the procedure, the predictor whose p-value of the corresponding t-statistic was highest is removed from the model. Once a regressor has been removed, it will be excluded from all further models. Removals continue until the candidate to exit has a p-value lower than 0.0000001. Our choices are motivated by the desire to limit the adverse effect of huge samples (which are typical in the STLF domain) on stepwise regression. See Lin et al. [21].

The regressors suspected to influence the current hourly load are

1. Loads from the last week: $L_{t-1}, L_{t-2}, \cdots, L_{t-170}$.
2. Hours of the day: 23 dummy variables excluding the 24-th hour;
3. Days of the week: 6 dummy variables excluding Wednesday;
4. Months of the year: 11 dummy variables excluding April;
5. Daylight saving time.
6. Public holidays (official and religious).
7. Local public holidays. Sardinia (April 8^{th}) e Sicily (May 15^{th}).
8. Holidays related to Easter.
9. Days near holidays.
10. Part-time holidays. These dummies indicate days traditionally considered public holidays, but eliminated in the revision of the holiday calendar: carnival (moving holiday), Saint Joseph (March 19^{th}), Saint Francis (October 4^{th}), alls ouls (November 2^{nd}), national unity day (November 4^{th})

Note that at least one category of each group of dummy variables must always be omitted to prevent complete collinearity.

The whole data set covers the period from May 5, 2013 to June 16, 2016 (a sample of 27′972 hourly observations). The same