Journal of the Korean Data & Information Science Society
2018, **29**(5), 1299–1307

http://dx.doi.org/10.7465/jkdi.2018.29.5.1299
한국데이터정보과학회지

Short term electricity demand forecasting in south korea with generalized additive models[†]

Hyelyen Kim[1] · Jaehee Kim[2]

[12]Department of Statistics, Duksung Women's University
Received 17 July 2018, revised 13 September 2018, accepted 17 September 2018

Abstract

Load forecasting is a key task for the effective operation and planning of power systems. This paper provides the Korean electricity load modeling and forecasting using SARMA (seasonal autoregressive moving average) models to explain an autoregressive lagged part and GAM (generalized additive model) to incorporate the explanatory variables. The selected exploratory variables are one-day-lagged loads, weather conditions, weekdays, and GDP as a global economic trend. The results demonstrate that our models are operationally efficient and achieve optimal prediction performance. The proposed model can explain the trend of Korean electricity demand and forecast the short-term demand.

Keywords: ARIMA, electricity demand forecasting, generalized additive model, load modeling, regression, smoothing methods, spline.

1. Introduction

The overall electricity consumption is a major measurement to indicate the degree of nations development since it might be a primary guideline for electricity system planning. The electricity is one of the major energy sources and the energy is the foundation of economic development. Brubacher and Wilson (1976) support that the electricity consumption forecast is especially important with regard to policy making in developing countries. Recently, for Korean electricity usage, Shin and Yoon (2016) provided accurate electricity demand forecasts with double seasonal Holt-Winters model and TBATS model in a specific time zone. Yoon and Choi (2015) studied to forecast the electricity demand using functional clustering according to weekdays or holidays.

We propose to use a semi-parametric approach using GAM that can carry out nonlinear effects and produce relatively parsimonious and interpretable models at the same time. Semi-parametric additive models were applied to long and short term load forecasting (Fan and Hyndman, 2012) for the Australian National Electricity Market, and for the French electricity load (Pierrot and Goude, 2011). We apply GAM methods on the Korean load consumption data, using the mgcv R package. Detailed explanations about GAM is provided with applications (Hastie and Tibshirani, 1990; Wood, 2006). We focus on modeling the Korean nationwide electricity load consumption.

[†] This research is supported by 2018 Duksung Women's University Research Fund.
[1] Undergraduate, Department of Statistics, Duksung Women's University, Seoul 01369, Korea.
[2] Corresponding author: Professor, Department of Statistics, Duksung Women's University, Seoul 01369, Korea. E-mail: jaehee@duksung.ac.kr

2. Electricity consumption data

We use Korean daily electricity consumption data obtained from January 2014 to December 2016. Figure 2.1 shows the raw data. The purpose of this research is to create a model to forecast the electricity consumption in nonlinear fashion.

As explanatory variables, we use daily meteorological data such as the temperature, the relative humidity, the sunshine duration, the cloud cover and the wind speed. The cloud cover is from 0 for no cloud to 10 for a very cloudy weather; the wind speed varies from 0 to 7 usually in Korea. These five national variables are computed over 50 stations in Korea, as a weighted mean. Figure 2.1 shows the seasonal patterns. Moreover, there are some important decreases in the load during the special national holidays such as Chuseok in fall and Lunar New Year Day in winter. We switched the national holiday values to the mean of the corresponding season. Figure 2.2 gives the weekday and weekend patterns. Each load is plotted together according to weekdays for three years. Notice that more load for weekdays due to working days and less for weekends.

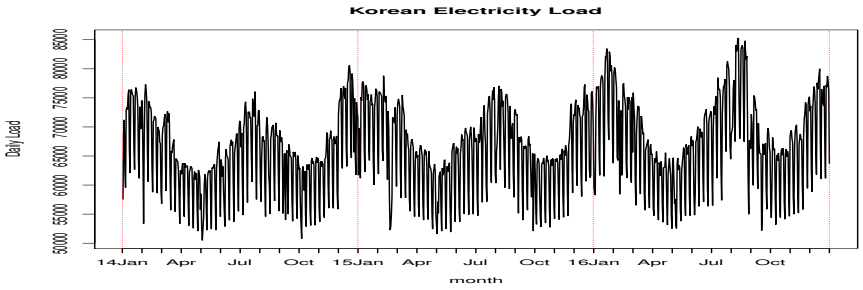

Figure 2.1 Korean daily electricity consumption data

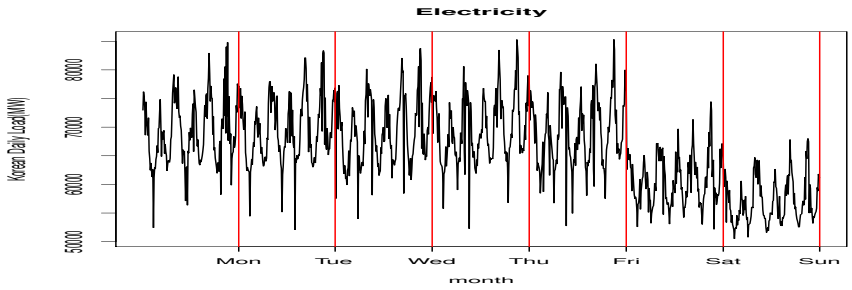

Figure 2.2 Electricity consumption by weekday

The many different effects in the Korean electricity demand make its modeling and forecasting a difficult task (Bruhns *et al.*, 2005). Those include a weekly seasonality, a growing trend, a dependence on the temperature and big differences between the summer and the winter due to the use of electrical air conditioning and heating.

3. Methodology

3.1. ARIMA and seasonal ARIMA

The autoregressive integrated moving average (ARIMA) representation is one of the popular linear models in time series forecasting during the past three decades (George and Gwilym, 1976). Time series forecasting is an important area of forecasting in which the past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship. The model is then used to extrapolate the time series into the future. This modeling approach is particularly useful when little knowledge is available on the underlying data generating process or when there is no satisfactory explanatory model that relates the prediction variable to other explanatory variables.

Let $\{Z_1, Z_2, \cdots, Z_n\}$ be a time series have the autoregressive and moving average, ARMA (p, q) process. This model is written as

$$Z_t = \mu + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}, \tag{3.1}$$

where $\{a_t\}$ is a white noise process which is iid with mean 0, and variance σ_a^2.

The homogeneous nonstationary sequences can be transformed into stationary sequences by taking successive differences of the series. The difference operator is denoted as $\bigtriangledown = 1 - B$ with the back operator B. When $\bigtriangledown^d Z_t \equiv W_t$ follows ARMA(p, q) model, Z_t is said to have ARIMA(p, d, q) model. That is, W_t follows ARMA(p, q) such as

$$W_t = \phi_1 W_{t-1} + \cdots + \phi_p W_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}. \tag{3.2}$$

Seasonality makes it so that the mean of the observations is not constant, but instead evolves according to a cyclical pattern. The most typical case is that we can incorporate seasonality into the ARIMA model multiplicatively, so that we obtain a multiplicative seasonal ARIMA model.

3.2. Generalized additive models with penalized regression splines

The additive model is a generalization of the usual linear regression model beyond the linearity. Additive models retain the important feature that the predictor effects are additive. With the added flexibility of non-parametric and additive regression models, there is always the risk of over-fitting the data and interpreting spurious features in the fitted curves. An additive model is defined as

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{3.3}$$

where the error ϵ's are independent of the X_j's with mean 0 and variance σ^2. The f_j's are arbitrary univariate functions, one for each predictor. In applications, it is useful to think of the additive model as a method for estimating simultaneously the appropriate metameter in which to measure the variables.

There are many ways to approach the formulation and estimation of additive models. The following conditional expectations provide a simple intuitive motivation for the back-fitting algorithm.

$$E[Y - \alpha - \sum_{j \neq k}^{p} f_j(X_j|X_k)] = f_k(X_k). \tag{3.4}$$

This suggests an iterative algorithm for computing all the f_j. The problem of estimating a generalized additive model becomes the problem of estimating smoothing parameters and model coefficients for a penalized likelihood maximization problem once a basis for the smooth functions has been chosen. This problem is solved by penalized iteratively reweighted least squares (P-IRLS). Thin-plate splines (Duchon, 1977) are a very elegant and general solution to the problem of estimating a smooth function of multiple predictor variables. Thin-plate spline smoothing estimates f by finding the function that minimizing

$$\|y - f\|^2 + \lambda J_{md}(f), \tag{3.5}$$

where y is the observed vector, $f = (f(x_1), \ldots, f(x_n))'$. $J_m d(g)$ is a penalty function measuring the wiggliness of f and λ is a smoothing parameter. For $2m > d$,

$$J_{md} = \int \cdots \int \sum_{v_1 + \cdots + v_d = m} \frac{m!}{v_1! \cdots v_d!} \left(\frac{\partial^m g}{\partial x_1^{v_1} \cdots \partial x_d^{v_d}}\right)^2 dx_1 \cdots dx_d. \tag{3.6}$$

In the P-IRLS method, which is implemented in the mgcv R package, the f_j are represented using regression splines. Given such bases, a GAM can be estimated as a GLM. To avoid overfitting, mgcv controls the smoothness for each term through a set of penalties applied to the likelihood of the GLM. Given a wiggliness measure of each function, a penalized likelihood for the model is fitted by iterative minimization in β, given λ_j, of the problem

$$l_p(\beta) = l(\beta) - \frac{1}{2} \sum_j \lambda_j \beta' S_j \beta, \tag{3.7}$$

where β contains θ and all the β_j, coordinates of f_j in its spline basis, S_j are matrices of known coefficients which penalize models with wiggly f_j. $l(\cdot)$ is the log-likelihood for β. The λ_j's are smoothing parameters that control the trade-off between fit and smoothness, and can be selected by minimization of the generalized cross validation (GCV) score.

4. Model comparison

Let Z_t denote the electricity load data. We use the log transformed data $Y_t = log(Z_t)$. Since DF (Dickey-Fuller) test for Y_t supports the stationarity. We go over the acf (autocorrelation function) for autocorrelations and pacf (partial autocorrelation function) plots for order and period information (Figure 4.1). The Seasonal model $SARIMA(1,0,1) \times (1,0,1)_7$ is chosen and its characteristic function is obtained as

$$(1 - 0.9281B)(1 - 0.9999B^7)y_t = (1 + 0.1934B)(1 + 0.9683B^7)\epsilon_t, \tag{4.1}$$

where ϵ_t's are white noises.

Table 4.1 provides the estimated parameters of the used Seasonal ARMA model. Figure 4.2 gives the acf and pacf (partial autocorrelation function) plots of $SARIMA(1,0,1) \times (1,0,1)_7$ residuals.

Table 4.1 $SARIMA(1,0,1) \times (1,0,1)_7$ Coefficients

	Estimated coeff	sd	Log likelihood	AIC
AR	0.9281	0.0132		
MA	-0.1934	0.0408	2086.26	4160.52
SAR	0.9999	0.0001		
SMA	-0.9683	0.0128		

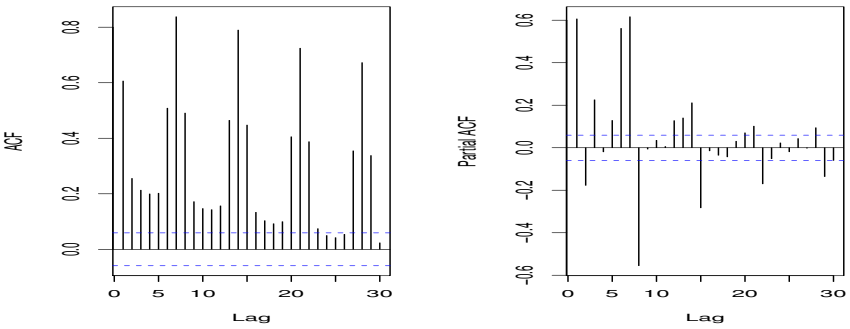

Figure 4.1 ACF and PACF plots of log transformed data

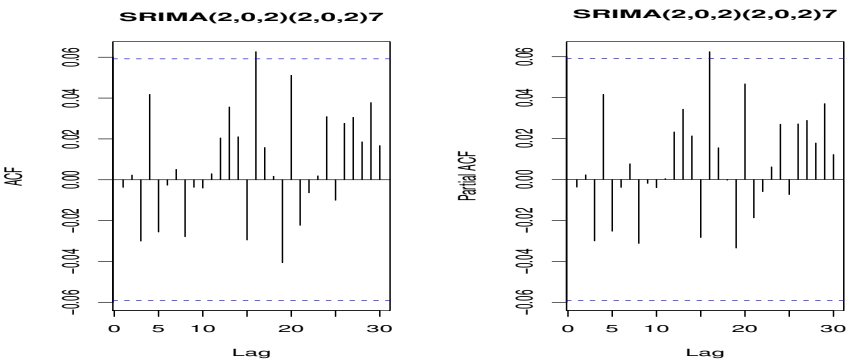

Figure 4.2 ACF and PACF plots of $SARIMA(1,0,1) \times (1,0,1)_7$

With the residuals from the proposed SARIMA model, GAM is fitted. Therefore SARIMA + GAM is estimated to incorporate the electricity-related covariates. Pierrot and Goude (2011) considered several covariates for GAM including weather factors and GDP and forecasted the electricity demand. Accordingly we take account several weather variables, GDP as a macro economic index and weekdays indicator as a related usage pattern variable.

For the covariate effects, GAM is fitted in the non-parametric and nonlinear fashion via smooth functions such as

$$f(\text{Temperature}) + f(\text{Cloud cover}) + f(\text{Relative humidity}) + \text{GDP} + \text{Week/weekend}. \quad (4.2)$$

We consider the model such as

$$\text{Model} = \text{SARMA} + \text{GAM} + \epsilon. \quad (4.3)$$

The Korean electricity load being strongly related to the current instant. Table 4.2 provides the different effects impacting the Korean demand, and the corresponding potential variables to model them.

The weekly seasonality is modeled using a categorical variable depending on the type of the day. For example, if the different levels are week day and weekend day, two levels of the load are estimated: one level for weeks and another level for weekends. Of course, this categorical type of the day variable may have different levels.

Table 4.2 Variable explanation

Effect	Variable in model
Economic growth	GDP
Weekly seasonality	Type of day: weekday vs weekend
Weather factor	Temperature of the current day
	Cloud cover
	Relative humidity

Our model is fitted over three years, from January 2014 to December 2016. Lunar Year Day and Chuseok (Korean Thanksgiving Day) can be considered as outliers for the normal trend. Therefore, we corrected the load on these special holidays with the average of the corresponding month and day when the holidays are included.

We checked for the residuals for model validity such that the normality hypothesis is not violated, there is no pattern left in the variance or in the mean, the residuals are well-scattered around 0, and there are no auto-correlations remaining. We choose one model over another according to the residuals and the mgcv measures of the fitting quality. The ER (error rate), MSE (mean squared error) and RMSE (root mean squared error) are useful in comparing models as follows:

$$\begin{aligned} \text{MSE} &= \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2, \\ \text{RMSE} &= \sqrt{MSE}, \\ \text{ER} &= 100 \times \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{y_t}(\%), \end{aligned}$$

where $\hat{y}$ is observed and $\hat{y}_t$ is the estimated value at time t.

We use thin plate regression spline bases: they can smooth any number of covariates, avoid putting knots and have some optimality properties. The choice of basis dimensions amounts to set the maximal possible degrees of freedom allowed for each model term. The number of effective degrees of freedom is estimated from data by GCV (generalized cross validation).

We introduce a new categorical variable in order to model the weekly seasonality: that is a week/weekends variable with two levels, a weekday level (=0) and a weekend level (=1). We compare the models from M1 to M5 in Table 4.3.

Figure 4.3 represents effects of exploratory variables in the GAM of model M4: the temperature effect, and the cloud cover effect. The Korean load demand strongly depends on the outside temperature. The colder the temperature is, the bigger the needs in electricity are, because of heating. There is also a slight cooling effect, when temperature is over 20. The lagged temperatures are significant because of the temperature reaction inertia. Finally, the power demand is growing when the weather is very cloudy (for a cloud cover bigger than 6). That can be explained by both a cold feeling amplified by clouds and bigger needs in lighting. Table 4.4 provides the GAM summary and Table 4.5 provides the linear term summary. We suggest M4 since M4 has the significant main effects of temperature and cloud cover. And the ER and RMSE of M4 is very close to those of M5 in Table 4.6.

Table 4.3 Model explanation

Model id	Variable to model
M1	SARMA
M2	M1+GAM(Temperature, Cloud cover, Relative humidity)
M3	M2+linear trend(GDP)
M4	M3+Week/weekends variable
M5	M4+interaction (Temperature, Cloud cover)

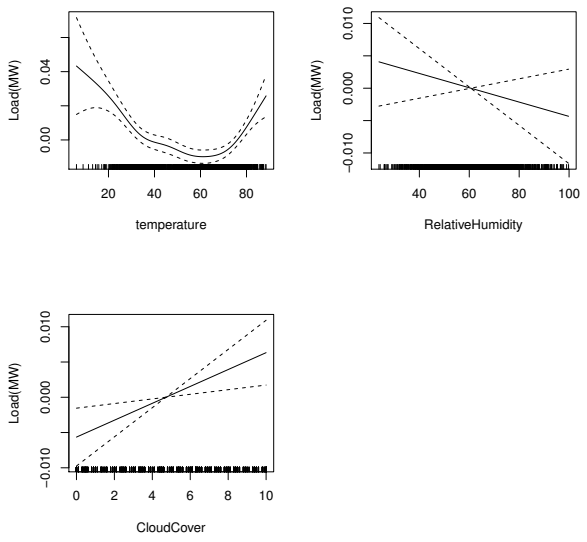

Figure 4.3 Components of GAM model fits

Figure 4.4 shows the predicted fits for the last 31 days and forecasts for 14 days ahead. We also plot the actual values which are provided beyond the observed days afterwards. The

Table 4.4 GAM summary

Covariates	Edf	F-stats	P-value
Temperature	1.809	16.897	< 0.00001
Relative humidity	1.000	1.421	0.23351
Cloud cover	1.000	7.583	0.00599

Table 4.5 Linear parameter coefficient estimates

Linear Variable	Estimate	Std.Error	t-stat	P-value
Intercept	-0.0129	0.0208	-0.621	0.535
GDP	3.6×10^{-8}	5.325×10^{-8}	0.678	0.498
Weekday/weekends	-0.0014	0.00229	-0.625	0.532

Table 4.6 Model comparison

	Predicted		Forecasting	
	ER	RMSE	ER	RMSE
M1	0.2083	0.0355	0.5198	0.0714
M2	0.1926	0.0341	0.3075	0.0415
M3	0.1923	0.0341	0.3079	0.0413
M4	0.1920	0.0341	0.3073	0.0413
M5	0.1893	0.0337	0.3324	0.0475

vertical line divides prediction and forecasting.

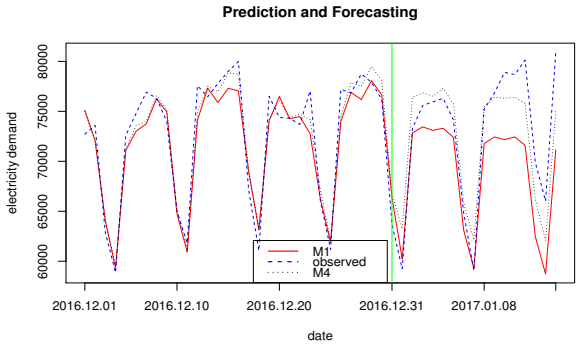

Figure 4.4 Prediction and forecasting plot

5. Conclusion

The study concludes, first, that the Korean electricity consumption is much affected by weekends, weather conditions and economic indicators. Second, the statistical models make better electricity demand projections, which may help the development of both a coherent energy policy in general and a healthy electricity market. Third, the statistical models with covariates will contribute to energy consumption pattern study, which will be useful to get

information about the electricity demand in Korea. The proposed model represents the time related pattern and includes the electricity demand related covariate effects. Forecasting can be improved by the suggested model by reducing forecasting error rates.

As a further study, the seasonal exponential smoothing model and the weighted support vector machines and others can be applied for the future model. The combined model (Bunn and Farmer, 1985) will be able to effectively map the seasonality and non-linearity which is normally presented in the electricity load data.

References

Brubacher, S. R. and Wilson, G. T. (1976). Interpolating time series with application to the estimation of holiday effects on electricity demand. *Applied Statistics*, 107-116.

Bruhns, A., Deurveilher, G. and Roy, J. S. (2005). A non linear regression model for mid-term load forecasting and improvements in seasonality. *Proceedings of the 15th Power Systems Computation Conference*, 22-26.

Bunn, D. W. and Farmer, E. D. (1985). *Comparative Models for Electrical Load Forecasting*, Wiley, New York.

Duchon, J. (1977). *Splines minimizing rotation-invariant semi-norms in Soboley spaces*. Constructive theory of functions of several variables, Springer, Berlin.

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

George, B. and Gwilym, J. (1976). *Time series analysis, forecasting and control*, Holden-Day, San Francisco.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, Chapman & Hall, London.

Pierrot, A. and Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. *Proceedings of ISAP Power*.

Shin, Y. and Yoon, S. (2016). Electricity forecasting model using specific time zone. *Journal of the Korean Data & Information Science Society*, **27**, 275-284.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*, Chapman & Hall, London.

Yoon, S and Choi, Y. (2015), Functional clustering for electricity demand data: A case study. *Journal of the Korean Data & Information Science Society*, **26**, 885-894.