International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:13, No:1, 2019

Forecasting Issues in Energy Markets within a Reg-ARIMA Framework

Ilaria Lucrezia Amerise

***Abstract*—**Electricity markets throughout the world have undergone substantial changes. Accurate, reliable, clear and comprehensible modeling and forecasting of different variables (loads and prices in the first instance) have achieved increasing importance. In this paper, we describe the actual state of the art focusing on reg-SARMA methods, which have proven to be flexible enough to accommodate the electricity price/load behavior satisfactory. More specifically, we will discuss: 1) The dichotomy between point and interval forecasts; 2) The difficult choice between stochastic (e.g. climatic variation) and non-deterministic predictors (e.g. calendar variables); 3) The confrontation between modelling a single aggregate time series or creating separated and potentially different models of sub-series. The noteworthy point that we would like to make it emerge is that prices and loads require different approaches that appear irreconcilable even though must be made reconcilable for the interests and activities of energy companies.

***Keywords*—**Forecasting problem, interval forecasts, time series, electricity prices, reg-plus-SARMA methods.

I. Introduction

A STRATEGIC objective of companies operating in energy markets is to have accurate forecasts that, by reducing uncertainty in predicting the effects of management, can lead to substantial improvements in efficiency in operations, reduction of maintaining costs and increased reliability of power supply.

Over the last few decades, there are many references to the efforts improving the accuracy of short term load forecasting. The classical linear regression is a popular tool to investigate the relationship between a set of regressors and load so as to forecast the load, on the basis of the values of the predictors. Model's parameters are estimated by applying the ordinary least squares technique, which involves the minimization of the sum of squared deviations (residuals) between observed expected values given the fitted model.

The goal of this paper is to discuss forecasting issues within the framework of the Reg-SARMA approach for short term forecasting of hourly electricity load. In the first stage, estimated loads are derived from a classical linear regression model (CLR) with non-stochastic predictors. In stage two, the residuals of stage one are examined by means of Box-Jenkins processes ([2]), to ascertain whether they are random, or whether they still bear patterns that can be used to improve fitting and enhance forecast accuracy. In this regard, we have ascertained ([1]) that the Reg-SARMA approach is not only effective in eliminating the harmful presence of serial dependence between regression residuals, but also easy to implement and satisfactory with respect of the predicted loads. In many respects, there remain some specific problems.

Ilaria L. Amerise (Dr.) is with the Department of Economics, Statistics and Finance, University of Calabria, Via P. Bucci, 87036, Rende (CS) Italy (e-mail: ilaria.amerise@unical.it).

The present paper proceeds as follows. After presenting the Reg-SARMA approach, Section II outlines the dichotomy between point and interval forecasts and analyzes the difficult choice between stochastic and non-deterministic regressors. Section III discusses thecomparison between modelling a unique hourly time series or creating separated and potentially different models (one of each hour) of multiple time series to be analyzed separately. Finally, Section IV contains concluding remarks and future research.

II. Reg-SARMA Approach

The proposed model has the following form

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

where L_t is the hourly electric price/load expressed in MWh and $X_{t,j}, j = 1, 2, \cdots, m$ are the regressors or predictors. Each β_j is a parameter that measures how Y_t is related to the j-th predictor: $\partial Y_t / \partial X_{t,j} = \beta_j$. Thus, coefficients measure the marginal effects of the predictor variables. One way to interpret β_0 is that it coincides with the expected price/load when all dummy variables are in their respective reference category. Each combination of $0/1$ gets its own regression surface, still parallel to each other. The addend e_t is an unobserved residual that accounts for disturbing factors other than the variation in the Y_t that predictors do not explain. we assume that the unobservable residuals follow a multiplicative SARMA process

$$e_t = \left[\phi^* (B)\right]^{-1} \theta^* (B)\, a_t \tag{2}$$

where B is the usual backward shift operator $B^j z_t = z_{t-j}$ and

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

are polynomials in B. The polynomials are constrained so that the roots of $\phi^*(B) = 0$ and $\theta^*(B) = 0$ have magnitudes strictly greater than one, with no single root common to both polynomials, that is, only processes which are stationary and invertible are considered. Because of the massive presence of binary variables in the regressors, the process in (2) does not include difference operators. The "burden of non-stationarity" is placed entirely on the orthogonal polynomials used as regressors. See [15].

Expression (2) may be considered as special case of the standard ARMA (p^*, q^*) by taking $p^* = p + sP$, $q^* = q + sQ$.

The integer s is the seasonal period. Note that some of the ϕ^*s and θ^*s could be put equal to zero. The errors a_ts are independent and identically distributed random variables with zero mean and finite variance σ_a^2.

The substitution of e_t in (1) with the process in (2) yields

$$Y_t = \beta_0 + \sum_{i=1}^{m} \beta_i X_{t,i} + \sum_{j=1}^{p^*} \phi^* e_{t-j} + \sum_{j=1}^{q^*} \theta^* a_{t-j} + a_t \ . \quad (4)$$

The estimated residuals can be written as

$$\widehat{e}_t = \left(Y_t - \widehat{Y}_t\right) = \sum_{j=1}^{p^*} \phi_j^* \widehat{e}_{t-j} - \sum_{j=1}^{q^*} \theta_j^* a_{t-j} + a_t, \quad t = 1, 2, \cdots, n \ . \quad (5)$$

Let us suppose for the moment that p^* and q^* are given. The estimation of the ARMA parameters can be carried out by optimizing the log-likelihood function of (5) provided that the a_ts were Gaussian random errors. Let $\widetilde{e}_t, t = 1, 2, \cdots, n$ be the estimates produced by the ARMA (p^*, q^*) process

$$\widetilde{e}_t = \sum_{j=1}^{p^*} \widehat{\phi}_j^* \widehat{e}_{t-j} - \sum_{j=1}^{q^*} \widehat{\theta}_j^* a_{t-j} + a_t, \quad t = 1, 2, \cdots, n \ . \quad (6)$$

The $\widetilde{e}_t$s can be substituted into (4) yielding

$$\widetilde{e}_t = \sum_{j=1}^{p^*} \widehat{\phi}_j^* \widetilde{e}_{t-j} - \sum_{j=1}^{q^*} \widehat{\theta}_j^* \widetilde{a}_{t-j}, \qquad t = 1, 2, \cdots, n \ . \quad (7)$$

where

$$\widetilde{a}_t = \begin{cases} 0 \ t = p^*, p^* - 1, \cdots \\ \bar{e}_t - \sum_{j=1}^{p^*} \widehat{\phi}_j^* \bar{e}_{t-j} - \sum_{j=1}^{q^*} \widehat{\theta}_j^* \widetilde{a}_{t-j} \ t = p^* + 1, p^* + 2, \cdots \end{cases} \quad (8)$$

In practice, what is done is to use the one-step-ahead forecast $\widetilde{e}_{t+1}$ as an estimate of the unknown e_t and set the unknown error a_{t+1} to its expected value of zero. The essence of the Reg-SARMA approach consists of the ordinary least squares applied to the estimation of the original model inclusive of the pseudo-regressors derived from residuals and errors observed at the first regression stage.

$$\widehat{Y}_t = \widehat{\beta_0} + \sum_{i=1}^{m} \widehat{\beta}_i X_{t,i} + \sum_{j=1}^{p^*} \widehat{\phi}_j^* \widehat{e}_{t-j} + \sum_{j=1}^{q^*} \widehat{\theta}_j^* \widetilde{a}_{t-j} \ . \quad (9)$$

Reg-SARMA equation (9) is a revised CLR model that should yield better statistics than the CLR model (1). See [11], [8] and [14]. One problem is still open. Since we ignore the orders of autoregressive-moving average components, the modelling procedure (5) should be repeated for each reasonable value of p^* and q^*. Let us assume that there is a true SARMA process for the time series: $(p^0, 0, q^0) \times (P^0, 0, Q^0)_s$ and fix the constraints $0 \leq p \leq \overline{p}, 0 \leq q \leq \overline{q}, 0 \leq P \leq \overline{P}, 0 \leq Q \leq \overline{Q}$, where $\overline{p}, \overline{q}, \overline{P}, \overline{P}$ are chosen beforehand trying to make sure the intervals include the true orders

$$p^0 \leq \overline{p}, \quad q^0 \leq \overline{q}, \quad P^0 \leq \overline{P}, \quad Q^0 \leq \overline{Q} \ . \quad (10)$$

One method used to locate a good solution is a trawling search through the $\overline{p} \cdot \overline{q} \cdot \overline{P} \cdot \overline{Q}$ possible processes. In general, brute-force methods are unmanageable for extremely long time series because of the computational complexity. If $\overline{p} = 4, \overline{q} = 4, \overline{P} = 3, \overline{Q} = 3$ then considering all possible processes involves estimating 400 different processes. Actually, the obstacle is more apparent than real. Improvements in computer technology and reductions in hardware costs allow us to consider the trawling search solution attractive for much more research and real-world applications than in the past.

Through extensive experimentation, [12] and [13] showed that GLS-type schemes allow the analyst to perform a generalized least squares estimation without the cumbersome computational difficulties associated with the inversion of large size variance-covariance matrices.

A. Point and Interval Forecasts

Energy companies are strongly affected by uncertain price/load conditions, as they are exposed to the different risks from liberalized energy markets in combination with important and, to a large extent, irreversible investments. Price/load predictions, however, are usually expressed as point forecasts that give little guidance as to their accuracy, whereas, the planning process needs to take into account the entire probability distribution of future prices/loads or at least intervals that have a pre-specified nominal coverage rate *i.e.* a given probability of containing the future prices/loads. The expanded relationship (9) can produce predictions of new values $\widehat{\mathbf{Y}}_{n,H} = (\widehat{Y}_{n+1}, \widehat{Y}_{n+2}, \cdots, \widehat{Y}_{n+H})$ given by

$$\widehat{\mathbf{Y}}_{n,H} = \mathbf{Z}_H \hat{\gamma} \text{ with } \mathbf{Z}_H = [\mathbf{X}_H | \mathbf{E}_H | \mathbf{A}_H], \ \hat{\gamma}^{\ t} = \left[\widehat{\boldsymbol{\beta}} | \widehat{\boldsymbol{\phi}} | \widehat{\boldsymbol{\theta}}\right] \quad (11)$$

where H is the number of prices/loads to be foreseen (lead time), $\mathbf{X}_H$ is a $[H \times (m+1)]$ matrix of the H predetermined values of the predictors, intercept included for $t = n, (n+1), \cdots, H$. $\mathbf{E}_H$ is a $(H \times p^*)$ matrix constructed by using the predicted values of the least squares residuals fitted by the selected SARMA process. Each column of $\mathbf{E}_H$ is a lagged instance of $\widetilde{e}_t$ at lags $1, 2, \cdots, p^*$ and $t = n, (n+1), \cdots, H$. Analogously, $\mathbf{A}_H$ is a $(H \times q^*)$ matrix constructed by using the estimated errors of selected SARMA process. Each column of $\mathbf{A}_H$ is a lagged instance of $\widetilde{a}_t$ at lags $1, 2, \cdots, q^*$ and $t = n, (n+1), \cdots, H$. The values of (11) serve to compute the diagnostic statistics for the new models. Load forecasting is necessary, but it is at least as important to provide an assessment of the uncertainty associated with forecasts. The usual method of evaluate the uncertainty associated with forecasts requires the computation of marginal prediction intervals at each individual horizon. However, Marginal intervals are overly optimistic, and may therefore be misleading since H marginal $100(1-\alpha)\%$ predictions give a probability lower than the nominal level $(1-\alpha)\%$ for the joint H intervals.

Managers of electric power and light systems are frequently confronted with decision problems that require assessing the set of possible upper/lower bounds that demand of electricity will follow over time and there are many well-known methods for computing simultaneous forecast intervals. Given the availability of H future values, a simple strategy is

to determine two bands such that, under the condition of independent Gaussian distributed random errors, the probability of consecutive future loads $L_{n+h}, h = 1, 2, \cdots, H$ lying simultaneously within their respective range is at least is $(1-\alpha)$

$$P\left[\bigcap_{h=1}^{H}\left(Y_{h,\alpha}^{1} \leq Y_{n+h} \leq Y_{h,\alpha}^{2}\right)\right] \geq (1-\alpha). \qquad (12)$$

where $Y_{h,\alpha}^{1} = Y_{n,h} - c_\alpha \sigma_h$, $Y_{h,\alpha}^{2} = Y_{n,h} + c_\alpha \sigma_h$ and σ_h is the standard deviation of the forecast error at the h-th time horizon. We propose the hyper-cuboid region

$$\left[\widehat{Y}_{n+h} \pm \widehat{\sigma}_h \; t\Big(\frac{\alpha}{2H}, n-\nu\Big)\right], \quad h = 1, 2, \cdots, H \; . \qquad (13)$$

where $\nu = m + 1$ is the number of estimated parameters, $c_\alpha = t(\alpha/2H, n-\nu)$ is the $\alpha/2H$ percentage point of the t distribution having $(n–\nu)$ degrees of freedom. Further, $\widehat{Y}_{n+h} = \widehat{\boldsymbol{\beta}}^{t} \mathbf{X}_{n+h}^{*}$. The estimate of σ_h is

$$\widehat{\sigma}_h = \widehat{\sigma}_e \sqrt{\frac{n+1}{n} + \sum_{i=1}^{m}\sum_{j=1}^{m} a^{i,j}\left(X_{h,i} - \bar{X}_i\right)\left(X_{h,i} - \bar{X}_i\right)} \; . \qquad (14)$$

Here $\widehat{\sigma}_e$ is the estimated mean square error of the regression and $\bar{X}_i$ is the mean of the i-th predictor. The quantity $a^{i,j}$ is the (i,j) element of the inverse matrix of the unscaled variance-covariance matrix of the predictors

$$a_{i,j} = \sum_{t=1}^{n} X_{t,i} X_{t,j} - n\bar{X}_i \bar{X}_j, \quad i, j = 1, 2, \cdots, m \qquad (15)$$

The intervals (13) give the box-shaped region in H-dimensional space that circumscribes the exact confidence ellipsoid of minimum volume. Forecasting the regression term would not present particular difficulties if regressors have a perfectly predictable nature. In the case of stochastic regressors, things change radically as several other sources of uncertainty must be taken into account.

B. Stochastic and Non-Deterministic Predictors

The preceding discussion assumes that the future values $\mathbf{Z}_H$ are known without errors or can be forecast perfectly or almost perfectly, ex ante. If, on the contrary, $\mathbf{Z}_H$ or part of it must themselves be forecast then formula (14) has to be modified to incorporate the uncertainty in forecasting the elements of $\mathbf{Z}_H$. Reference [10] [Section 4.6.4] observes that firm analytical results for the correct forecast variance for this case remain to be derived except for simple special cases. For example, formula (14) may lead to a serious underestimate of the forecast standard error, if, as is not uncommon in practice, the uncertainty about the future value of predictors is of the same order of magnitude as the uncertainty about the regression residuals.

To keep the problem of estimating (1) tractable, we may use deterministic exogenous variables so we know exactly what they will be at any future time (*e.g.* calendar variables, polynomials or sinusoids in time). This choice is also suggested by the fact that non-deterministic exogenous regressors, which must also be forecast, is one of the possible causes of inefficiency in prediction intervals. See [3] [Section 6.5].

Predictors such as sociological and demographic factors, temperature, relative humidity, solar radiation, wind speed, cloud cover, etc. should be ignored because they are unusable in aggregation at large regional scale. Additionally, prediction of changes in social and climate factors raises the question of whether these variables are predictable and, if they are, whether predictability can be achieved for changes at an hourly time step (see, for example, [6] and [4]).

Additionally, if one or more predictors must, themself, be forecast, then the formula for forecast variances would have to be modified to incorporate the uncertainty in forecasting those predictors that are not known perfectly, ex ante. This will vastly complicate the computation, in particular for taking into account relationships between the errors in the process generating the predictors and the errors in the process generating the loads (see [7]). Many authors view this problem as simply intractable. See [10] [p. 127-128]. In any event, indicators for the climate could be introduced at the cost of setting load forecasting into a more general framework of a system of time series regression equations (some based on atmospheric physics) that are outside the scope of this discussion. The undeniable influences of climatic variations are captured implicitly by the joint action of a polynomial trend and calendar dummies. The selected SARMA process serves to compute, standing at time n, forecast $\widehat{Y}_{n+k}$ of the price at day $n+k, k = 1, 2, \cdots, H$ which are optimum in the sense of quadratic loss, conditional on an information set $I_n = \{Y_1, Y_2, \cdots, Y_n\}$, i.e. $\widehat{Y}_{n+k} = E(Y_{n+k}|I_n), k = 1, 2, \cdots, H$. It turns out that, under reasonably weak conditions, the optimal forecast is the expected value of the series being forecast, conditional on available information. See [5] [p. 172].

III. Confrontation between Single or Multiple Time Series Models

The prediction of hourly electricity demand follows two directions of research. First, the 24 observations are combined to make a single consecutive time series. In contrast, given the large amount of data which is generally available, it is possible the treating of each hour as a separate time series such that 24 different models are estimated. Of course, the same applies for the different hourly time bands. One defect that emerges from the use of a single time series is that we work on long time series ($27'972$ hourly observations) See for example [15]. On the other hand, dealing with, for example daily time series (one for each hour) reduces the length of the sequences, but the recency effect is attenuate. Many studies state that produces and consumes remember the past and this implies that the electricity price/load in previous period will matter in decision today.

Research have to decide whether to use the data as a single consecutive sequence or to develop a separate model for different hours of the day or different time-bands. One of the advantages of working with a single time series is that it exploits the correlation between hourly trends. For what concerns the loads the presence of serial correlation reveals

that there is additional information in the data that has not been exploited in the classical linear regression model. See [9] [Ch. 17]. This is a fact that of which we are fully aware as in model (1) we have omitted to account for short-run effects on electricity demand. It follows that, given the heavy serial correlation in the data, we need to find a way to incorporate that information into the regression model itself. With regard to prices, offers are made on groups of hours that imply correlation between hours. For a number of reasons, global models are not satisfactory. For instance, they may require modeling complicated intra-day patterns in the hourly sequences. On the other hand, an hour by hour modeling may need to estimate too many parameters. If we work on multiple time series, the model is simplified and it becomes sensitive to the hourly rate, moreover, the presence of dummies is reduced. For the purposes of the prediction of the day-head it becomes an aggregate of the forecast of the hours. On the basis of these considerations, it is reasonable to conclude that an effective analysis for short-term price/load forecasting should consider splitting the total time series into sub-series, to be modelled separately from other time-bands.

IV. Conclusion

We have worked with hourly time series both of prices and loads preferring a one model formulation for the latter and a multimodel formulation for the former. This view seems to be much more in line with the functioning of the electricity market in which the demand for and supply of electricity follows different paths. However, prices and loads do not necessary require different statistical approaches; actually, a common model can favor deeper and more careful analysis of the relationships between production and consumption. We are convinced that even though the models usually described in the literature are irreconcilable, they must be made reconcilable for the interests and activities of energy companies. As a future research, we plan to study a reg-VARMA model that contains both loads and prices.

References

[1] Amerise, I. L., Tarsitano, A.: Point and interval forecasts of electricity demand with Reg-SARMA models. Submitted (2018).
[2] Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco (1976).
[3] Chatfield, C. (2000). *Time series forecasting*. Chapman & Hall/CRC, Boca Raton.
[4] Charlton, N., 1, Singleton, C.: A refined parametric model for short term load forecasting. International Journal of Forecasting, 30 364–368 (2014).
[5] Diebold, F. X. (2007). *Elements of forecasting. 4th Edition*. Thomson South-Western. Available on line: http://threeplusone.com/fieldguide.
[6] Engle, R. and Mustafa, C. and Rice, J. (1992). "Modeling peak electricity demand". *Journal of Forecasting*, 11, 241 – 251.
[7] Feldstein, M. S.: The error of forecast in econometric models when the forecast-period exogenous variables are stochastic. Econometrica, 39, 55–60 (1971).
[8] Findley, D. F., C. Monsell, B. C., Bell, W. R., Otto, M. C., Chen, B-C.: An iterative GLS approach to maximum likelihood estimation of regression models with ARIMA errors. Journal of Business & Economic Statistics, 16, 127–152 (1998).
[9] Gilchrist, W.: Statistical Forecasting. John Wiley & Sons, London (1976).
[10] Green, W. H.: Econometric Analysis (7th Edition): International edition. Pearson Education Limited (2012).
[11] Harvey, A. C., Phillips, G. D. A.: Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. Biometrika, 66, 49–58 (1979).
[12] Koreisha, S. G., Pukkila, T.: Linear methods for estimating ARMA and regression models with serial correlation. Communications in Statistics-Simulation, 19, 71–102 (1990).
[13] Kavalieris, L., Hannan, E. J., Salau, M.: Generalized least squares estimation of ARMA models. Journal of Time Series Analysis, 24, 165–172 (2003).
[14] Poskitt, D., Salau, M.: On the relationship between generalized least squares and Gaussian estimation of vector ARMA models. Journal of Time Series Analysis, 16, 617–645 (1995).
[15] Tarsitano, A., Amerise, I. L.: Short-term load forecasting using a two-stage sarimax model. Energy, 133, 108–114 (2017).