Advanced Engineering Informatics 35 (2018) 1–16

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Full length article

Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia

Mohanad S. Al-Musaylh[a,b,*], Ravinesh C. Deo[a,d,*], Jan F. Adamowski[c], Yan Li[a]

[a] School of Agricultural, Computational and Environmental Sciences, Institute of Agriculture and Environment (IAg&E), University of Southern Queensland, QLD 4350, Australia

[b] Management Technical College, Southern Technical University, Basrah, Iraq

[c] Department of Bioresource Engineering, Faculty of Agricultural and Environmental Science, McGill University, Québec H9X 3V9, Canada

[d] Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou, China

ARTICLE INFO

Keywords:
Electricity demand forecasting
Machine learning
SVR
MARS
ARIMA

ABSTRACT

Accurate and reliable forecasting models for electricity demand (*G*) are critical in engineering applications. They assist renewable and conventional energy engineers, electricity providers, end-users, and government entities in addressing energy sustainability challenges for the National Electricity Market (NEM) in Australia, including the expansion of distribution networks, energy pricing, and policy development. In this study, data-driven techniques for forecasting short-term (24-h) *G*-data are adopted using 0.5 h, 1.0 h, and 24 h forecasting horizons. These techniques are based on the Multivariate Adaptive Regression Spline (MARS), Support Vector Regression (SVR), and Autoregressive Integrated Moving Average (ARIMA) models. This study is focused in Queensland, Australia's second largest state, where end-user demand for energy continues to increase. To determine the MARS and SVR model inputs, the partial autocorrelation function is applied to historical (area aggregated) *G* data in the training period to discriminate the significant (lagged) inputs. On the other hand, single input G data is used to develop the univariate ARIMA model. The predictors are based on statistically significant lagged inputs and partitioned into training (80%) and testing (20%) subsets to construct the forecasting models. The accuracy of the *G* forecasts, with respect to the measured *G* data, is assessed using statistical metrics such as the Pearson Product-Moment Correlation coefficient (*r*), Root Mean Square Error (*RMSE*), and Mean Absolute Error (*MAE*). Normalized model assessment metrics based on *RMSE* and *MAE* relative to observed means ($RMSE_{\overline{G}}$ and $MAE_{\overline{G}}$), Willmott's Index (*WI*), Legates and McCabe Index (E_{LM}), and Nash–Sutcliffe coefficients (E_{NS}) are also utilised to assess the models' preciseness. For the 0.5 h and 1.0 h short-term forecasting horizons, the MARS model outperforms the SVR and ARIMA models displaying the largest *WI* (0.993 and 0.990) and lowest *MAE* (45.363 and 86.502 MW), respectively. In contrast, the SVR model is superior to the MARS and ARIMA models for the daily (24 h) forecasting horizon demonstrating a greater *WI* (0.890) and *MAE* (162.363 MW). Therefore, the MARS and SVR models can be considered more suitable for short-term *G* forecasting in Queensland, Australia, when compared to the ARIMA model. Accordingly, they are useful scientific tools for further exploration of real-time electricity demand data forecasting.

Abbreviations: MW, Megawatt; *G*, Electricity load (demand, Mega Watts); MARS, Multivariate Adaptive Regression Splines; SVR, Support Vector Regression; ARIMA, Autoregressive Integrated Moving Average; *r*, Correlation Coefficient; *RMSE*, Root Mean Square Error (MW); *MAE*, Mean Absolute Error (MW); $RMSE_{\overline{G}}$, Relative Root Mean Square Error, %; $MAE_{\overline{G}}$, Mean Absolute Percentage Error, %; *WI*, Willmott's Index of Agreement; E_{NS}, Nash–Sutcliffe Coefficient; E_{LM}, Legates and McCabe Index; ANN, Artificial Neural Network; RBF, Radial Basis Function for SVR; σ, Kernel Width for SVR Model; C, Regulation for SVR Model; $BF_m(X)$, Spline Basis Function for MARS; *GCV*, Generalized Cross-Validation; *p*, Autoregressive Term in ARIMA; *D*, Degree of Differencing in ARIMA; *Q*, Moving Average Term in ARIMA; AEMO, Australian Energy Market Operator; NEM, National Electricity Market; ACF, Auto-Correlation Function; PACF, Partial Auto-Correlation Function; *MSE*, Mean Square Error (MW); R^2, Coefficient of Determination; *AIC*, Akaike Information Criterion; *L*, Log Likelihood; σ^2, Variance; G_i^{for}, i^{th} Forecasted Value of *G*, MW); G_i^{obs}, i^{th} Observed Value of *G*, MW); Q_{25}, Lower Quartile (25th Percentile); Q_{50}, Median Quartile (50th Percentile); Q_{75}, Upper Quartile (75th Percentile); *d*, Degree of Differencing in ARIMA

* Corresponding authors at: School of Agricultural, Computational and Environmental Sciences, Institute of Agriculture and Environment (IAg&E), University of Southern Queensland, QLD 4350, Australia.

E-mail addresses: MohanadShakirKhalid.AL-Musaylh@usq.edu.au, mohanadk21@gmail.com (M.S. Al-Musaylh), ravinesh.deo@usq.edu.au (R.C. Deo).

https://doi.org/10.1016/j.aei.2017.11.002
Received 8 April 2017; Received in revised form 18 November 2017; Accepted 20 November 2017
1474-0346/

1. Introduction

Electricity load forecasting (also referred to as demand and abbreviated as *G* in this paper, MW) plays an important role in the design of power distribution systems [1,2]. Forecast models are essential for the operation of energy utilities as they influence load switching and power grid management decisions in response to changes in consumers' needs [3]. *G* forecasts are also valuable for institutions related to the fields of energy generation, transmission, and marketing. The precision of *G* estimates is critical since a 1% rise in load forecasting error can lead to a loss of millions of dollars [4–6]. Over- or under-projections of *G* can endanger the development of coherent energy policies and hinder the sustainable operation of a healthy energy market [7]. Furthermore, demographic, climatic, social, recreational, and seasonal factors can impact the accuracy of *G* estimates [1,8,9]. Therefore, robust forecasting models that can address engineering challenges, such as minimizing predictive inaccuracy in *G* data forecasting, are needed to, for example, support the sustainable operation of the National Electricity Market (NEM).

Qualitative and quantitative decision-support tools have been useful in *G* forecasting. Qualitative techniques, including the Delphi curve fitting method and other technological comparisons [6,10,11], accumulate experience in terms of real energy usage to achieve a consensus from different disciplines regarding future demand. On the other hand, quantitative energy forecasting is often applied through physics-based and data-driven (or black box) models that draw upon the inputs related to the antecedent changes in *G* data. The models' significant computational power has led to a rise in their adoption [12]. Data-driven models, in particular, have the ability to accurately forecast *G,* which is considered a challenging task [6]. Having achieved a significant level of accuracy, data-driven models have been widely adopted in energy demand forecasting (*e.g.*, [13,14]). Autoregressive Integrated Moving Average (ARIMA) [15], Artificial Neural Network (ANN) [16], Support Vector Regression (SVR) [17], genetic algorithms, fuzzy logic, knowledge-based expert systems [18], and Multivariate Adaptive Regression Splines (MARS) [19] are among the popular forecasting tools used by energy researchers.

The SVR model, utilised as a primary model in this study, is governed by regularization networks for feature extraction. The SVR model does not require iterative tuning of model parameters [20,21]. Its algorithm is based on the structural risk minimization (SRM) principle and aims to reduce overfitting data by minimizing the expected error of a learning machine [21]. In the last decades, this technique has been recognized and applied throughout engineering, including in forecasting (or regression analysis), decision-making (or classification works) processes and real-life engineering problems [22]. Additionally, the SVR models have been shown to be powerful tools when a time-series (*e.g.*, *G*) needs to be forecasted using a matrix of multiple predictors. As a result, their applications have continued to grow in the energy forecasting field. For example, in Turkey (Istanbul), several investigators have used the SVR model with a radial Basis Kernel Function (RBF) to forecast *G* data [23]. In eastern Saudi Arabia, the SVR model generated more accurate hourly *G* forecasts than a baseline autoregressive (AR) model [24]. In addition, different SVR models were applied by Sivapragasam and Liong [25] in Taiwan to forecast daily loads in high, medium, and low regions. In their study, the SVR model provided better predictive performance than an ANN approach for forecasting regional electric loads [29]. Except for one study that confirmed SVR models' ability to forecast global solar radiation [17], to the best of the authors' knowledge, a robust SVR forecasting model has been limitedly applied for energy demand. Thus, additional studies are needed to explore SVR modelling in comparison to other models applied in *G* forecasting.

Contrary to the SVR model, the MARS model has not been widely tested for *G* forecasting. It is designed to adopt piecewise (linear or cubic) basis functions [26,27]. In general, the model is a fast and flexible statistical tool that operates through an integrated linear and non-linear modelling approach [28]. More importantly, it has the capability of employing a set of basic functions using several predictor variables to assess their relationship with the objective variable through non-linear and multi-collinear analysis. This is important for demand forecasting based on interactions between different variables and the demand data. Although the literature on MARS models applied in the field of *G* forecasting is very scarce, this model has proven to be highly accurate in several estimation engineering challenges. Examples may be drawn from studies that discuss doweled pavement performance modelling, determination of ultimate capacity of driven piles in cohesionless soil, and analysis of geotechnical engineering systems [29–31]. In Ontario (Canada), the MARS model was applied, through a semiparametric approach, for forecasting short-term oil prices [32] and investigating the behaviour of short-term (hourly) energy price (HOEP) data through lagged input combinations [8]. Sigauke and Chikobvu [19] tested the MARS model for *G* forecasting in South Africa; this demonstrated its capability of yielding a significantly lower Root Mean Square Error (*RMSE*) when compared to piecewise regression-based models. However, despite its growing global applicability (*e.g.*, [26,27,33–35]), the MARS model remains to be explored for *G* forecasting in the present study region.

In the literature, the ARIMA model has generated satisfactory results for engineering challenges including the forecasting of electricity load data [15], oil [32], and gas demand [36]. A study in Turkey applied a co-integration method with an ARIMA model for *G*-estimation and compared results with official projections. It concluded that approximately 34% of the load was overestimated when compared to measured data from the ARIMA model [8]. Several studies have indicated that the ARIMA model tends to generate large errors for long-range forecasting horizons. For example, a comparison of the ARIMA model, the hybrid Grey Model (GM-ARIMA), and the Grey Model (GM(1, 1)) for forecasting *G* in China showed that GM (1, 1) outperformed the ARIMA model [37]. Similarly, a univariate ARAR model (i.e., a modified version of the ARIMA model) outperformed a classical ARIMA model in Malaysia [38]. However, to the best of the authors' knowledge, a comparison of the MARS, SVR, and ARIMA methods, each having their own merits and weaknesses, has not been undertaken in the field of G forecasting.

To explore opportunities in *G* forecasting, this paper discusses the versatility of data-driven techniques (multivariate MARS and SVR models and the univariate ARIMA model) for short-term half-hourly (0.5 h), hourly (1.0 h) and daily (24 h) horizon data. The study is beneficial to the field of power systems engineering and management since energy usage in Queensland continues to face significant challenges, particularly as it represents a large fraction (i.e., 23%) of the national 2012–2013 averaged energy demand [39]. The objectives of the study are as follows: (1) To develop and optimise the MARS, SVR, and ARIMA models for *G* forecasting using lagged combinations of the state-aggregated *G* data as the predictor variable; (2) To validate the optimal MARS, SVR, and ARIMA models for their ability to generate *G* forecasts at multiple forecasting horizons (i.e., 0.5, 1.0 and 24 h); and (3) To evaluate the models' preciseness over a recent period, [01-01-2012 to 31-12-2015 (dd-mm-yyyy)], by employing robust statistical metrics comparing forecasted and observed *G* data obtained from the Australian Energy Market Operator (AEMO) [40]. To evaluate and reach these objectives, this paper is divided into the following sections: Section 2 describes the theory of SVR, MARS, and ARIMA models; Section 3 presents the materials and methods including the *G* data and model development and evaluation; Section 4 presents the results and discussion; and Section 5 further discusses the results, research opportunities, and limitations. The final section summarizes the research findings and key considerations for future work.

2. Theoretical background

2.1. Support Vector regression

An SVR model can provide solutions to regression problems with multiple predictors $X = \{x_i\}_{i=1}^{i=n}$, where n is the number of predictor