
Proceedings Paper:
Scott, D., Simpson, T., Dervilis, N. et al. (2 more authors) (2018) Machine learning for energy load forecasting. In: Journal of Physics: Conference Series. Modern Practice in Stress and Vibration Analysis (MPSVA) 2018, 02-04 Jul 2018, Cambridge, United Kingdom. IOP Publishing .

https://doi.org/10.1088/1742-6596/1106/1/012005

Machine Learning for Energy Load Forecasting

IOP Conf. Series: Journal of Physics: Conf. Series **1106** (2018) 012005 doi:10.1088/1742-6596/1106/1/012005

Machine Learning for Energy Load Forecasting

David Scott[1], Tom Simpson[1], Nikolaos Dervilis[2], Timothy Rogers[2], Keith Worden[2],

[1] Department of Mechanical Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK
[2] Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK

E-mail: n.dervilis@sheffield.ac.uk

Abstract. With an increasing penetration of renewables into energy markets, it is desirable to have a flexible grid in order to match large fluctuations in supply to a volatile power output typical of renewable supply. Hence, it is imperative to accurately forecast power load demand. The recent emergence of big data analytics and machine learning techniques have shown great success in a wide range of regression problems in varied industries and various data can be harnessed by the energy industry to better understand likely energy loads placed upon the system. This paper presents a comparison of several regression models which can be used for accurate predictions of energy load given environmental feature data. Here we show that dynamic Gaussian Processes can be used as a powerful tool taking into account the non-stationarity of the data under analysis. This regression model was compared Neural Networks, used most extensively in the industry, and linear regression models to give an idea of their comparable accuracy. However, it was noted that the dynamic Gaussian Process were inferior to a Neural Network when training for huge datasets due to their high relative computational cost, increased uncertainty with projection time, and large memory usage. Though primarily used for dynamics problems, there are a range of non-stationary problems which could benefit from the use of a dynamic Gaussian Process of which this paper just presents one. It also considers online learning models be used for real time forecasting.

1. Introduction

1.1. Energy Load Forecasting

Accurate prediction of grid electricity demand is key to the operation of a power utility company. Load forecasting is essential to ensure that sufficient capacity is available to maintain system frequency whilst minimising the amount of redundant generation and the use of backup peaker plants [1]. The increasing penetration of variable renewable energy sources in utility generation is resulting in an increased requirement for grid flexibility and accurate predictives, especially if the eventual goal is the complete phasing out of fossil fuel generation [2].

Load forecasting falls into two distinct categories: long term forecasting and short/medium term forecasting [1]. Long term forecasting is principally interested in predicting multi-year trends in electricity demand which are to influence decisions regarding infrastructure investment. This requires a lot of contextual information regarding governmental policy, economic growth and industrial activity etc. within the country of interest. Conversely, short-medium term forecasting focuses on the prediction of power demand on a day to day basis in order to inform the management of operational assets. This short/medium term forecasting is of greater importance

when it comes to maximising the performance of renewable energy assets and is the principal interest of this paper.

1.2. Machine Learning

Machine learning is the name for a set of mathematical techniques which can be applied to data to find patterns within it. Advances in computers have allowed these computationally expensive methods to be used on vast datasets where patterns may be obscured by the noise or scale of the data. Patterns can be used to form models to make predictions on new data.

In supervised learning, one is given $(x_1, y_1), (x_2, y_2)(x_n, y_n)$ and must find the relationship $f(x) = y$ which takes a new datapoint $\mathbf{x}_i$ and produces a prediction of y_i [3]. In regression, y_i is a point whilst in classification problems, it is a class. In this paper, regression techniques are used to predict the energy load of an area in the UK given environmental data and other information relevant to energy use such as school holidays.

2. Models

2.1. Linear Regressions

A linear regression model fits a function to data using the squared euclidean distance metric to optimise the functions fit to training data. In the most simple model, $y = \beta_1 x + \beta_0$, a straight line is fitted to the data with gradient β_1 and y-intercept β_0. With larger numbers of features, this becomes a hyperplane.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{1}$$

Interaction terms $x_i x_j$ can be included to map the relationship between terms x_i and x_j. This yields better results than a simple linear regression if the effect of x_i on y depends on the other predictors $x_{j,j\neq i}$.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{12} x_1 x_2 + \ldots + \beta_{1k} x_1 x_k + \ldots + \beta_{(k-1)k} x_{k-1} x_k \tag{2}$$

In a quadratic model, a quadratic function is fit to the data and optimised in a least squares manner. Consequently, the model contains an intercept, linear terms, interactions, and squared terms as visualised in Equation 3.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \beta_{12} x_1 x_2 + \ldots + \beta_{1k} x_1 x_k + \ldots + \beta_{(k-1)k} x_{k-1} x_k + \beta_{11} x_1^2 + \ldots + \beta_{kk} x_k^2 \tag{3}$$

2.2. Gaussian Processes

Gaussian process regression is a non-parametric regression technique widely used for approximating arbitrary non-linear functions. It has several desirable characteristics that make it rather suitable for the problem studied herein. Firstly, being non-parametric it is not required that any structure be imposed on the data as is necessary in polynomial regression with regards to the number of terms considered or in neural-networks in which network architecture must be selected. Secondly, Gaussian process regression is derived purely from well known rules for combining normal distributions using Bayes theorem. As a result it naturally provides error bounds for the predictions made by the algorithm, for any practical application the error bounds are likely almost equally important as the accuracy of the prediction itself.

A Gaussian process regression predicts a probability distribution for a given input vector based on previous training input vectors and associated responses. The predicted distribution is given by Equation 4 as derived in detail by Rasmussen [4]. That is the probability distribution of the predictor function f_s at points of interest X_s given the training values of X and y, is a Gaussian distribution with mean vector $\sigma_n^{-2} X_s^T A^{-1} X y$ and covariance matrix given by

$X_s^T A^{-1} X_s$. In these equations σ_n is the assumed standard deviation of the noise in the measurement values and Σ_p is the covariance function.

$$P(f_s|X_s, X, y) = \mathcal{N}(\sigma_n^{-2} X_s^T A^{-1} X y, X_s^T A^{-1} X_s) \tag{4}$$

$$A = \sigma_n^{-2} X X^T + \Sigma_p^{-1} \tag{5}$$

Gaussian process regression models can also be naturally extended to take previous function values as input known as auto-regressive exogenous Gaussian process models. Further details on auto-regressive exogenous models are given in Section 2.4.

2.3. Neural Networks

Loosely based on the brain, the neural network describes an iterative procedure by which a collection of units known as neurons transmit and process signals in order to create an associative mapping of $X \mapsto Y$. The method has become popular with the advent of huge quantities of computational power which have enabled large combinations of equations to be solved quickly.

Each node is a non-linear activation function, Equation 6, each of which are arranged in layers through which a signal moves through an input layer, passes through various hidden layers, and finally is transmitted to an output layer. The activation function describes the movement from layer j to layer i with a weight w_{ij} which are summed at node i to be passed through the activation function.

$$x_i = f\left(\sum_j w_{ij} x_j\right) \tag{6}$$

A learning procedure is carried out to optimise the mapping by changing the weights at each neuron to best describe the general mapping of $X \mapsto Y$ without overfitting. There are various methods of doing this; the most simple of which is Gradient Descent. After a training procedure, the neural network should accurately predict new y values given new x values. More details about the process of neural networks can be found in [3].

2.4. Auto-regressive with Exogenous Variables Models

An auto-regressive model is a regression model which takes as input the previous n values of the system response, where n is the order of the model, and predicts the system value at the next time step as a function of these previous responses. These models can be very powerful in the case of dynamic systems in which current and previous system states can have considerable impact on future values.

$$y(t_i) = f(y(t_{i-1}), y(t_{i-2}), \ldots, y(t_{i-n})) \tag{7}$$

Auto-regressive, exogenous models combine a more traditional regression model with auto-regressive models. These models use both normal input vectors, for example in this case: the temperature data, and previous predictions made by the model.

$$y(t_i) = f(x_1, x_2, \ldots, x_m, y(t_{i-1}), y(t_{i-2}), \ldots, y(t_{i-n})) \tag{8}$$

A key parameter of auto-regressive models which must be stipulated is the number of previous model predictions taken as inputs, the order of the model. This must be done with care as similarly to most regression algorithms there is a trade-off between model flexibility and over-fitting of the training data.

3. Methods

3.1. Preprocessing

The data were scaled to have a zero mean and unit variance to ensure that scaling effects of data with different magnitudes did not affect regression performance. Categorical data were converted to numerical data and the date information was separated into three features indicating day of the week, day of the year, and year reflecting the seasonality of the data. Analysis showed that this separation doubled performance accuracy when using a dynamic Gaussian Process $n = 7$, reducing the prediction performance error from 0.45% to 0.26%.

3.2. Optimisation of NARX Gaussian Process

An investigation into the optimal number of previous time steps used was performed on the data varying n between 1 and 10, visualised in Figure 1. It was found that 7 time points was the optimal number for the dataset under investigation.

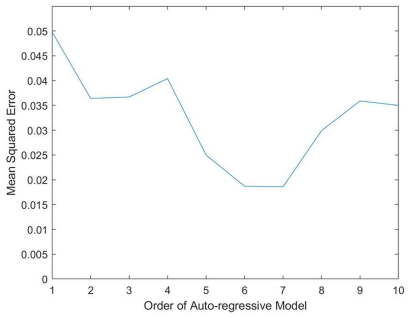

Figure 1. Optimisation of number of previous time steps used in model.

3.3. Optimisation of Neural Network

The neural network's shape was difficult to optimise given the high variability of the function to fall into local minima given random initial weights. To mitigate these circumstances, a Bayesian neural network was used with a single layer of 10 neurons.

4. Results

Table 1. Results of different regression methods on energy load data.

Method	Accuracy (%)
Linear Regression	13.78
Linear Regression with Interactions	10.80
Linear Regression (Quadratic)	6.36
Neural Network	28.15
NARX Gaussian Process	0.26

It is clear from this table that the NARX Gaussian process significantly outperformed all other algorithms trained on the training data due to its ability to use the previous states as predictors for future states. A detailed examination of the performance of all the algorithms is given in the following sections.

4.1. Linear Regression

The three linear regression models increased in accuracy as their complexity increased, a predictable result given the clear non-linearities in the dataset.

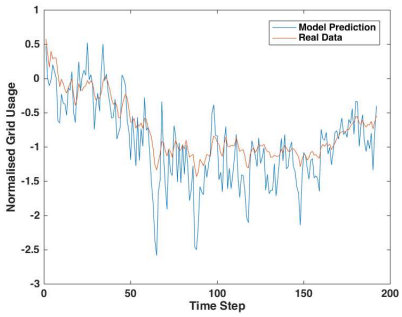

Figure 2. Linear regression.

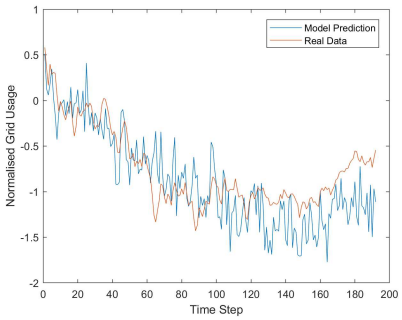

Figure 3. Linear regression with interaction terms.

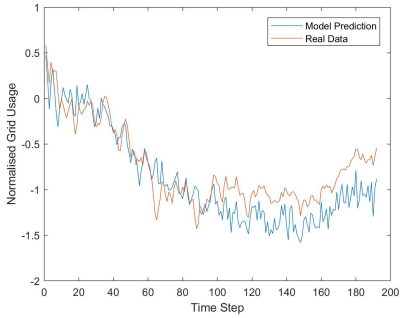

Figure 4. Linear regression (quadratic).

4.2. Neural Network

The neural network had a large degree of variability due to random initial weights as shown by Figure 6. With a larger training data set, the network could likely generate more accurate predictions. Using an auto-regressive network would also be likely to boost performance.

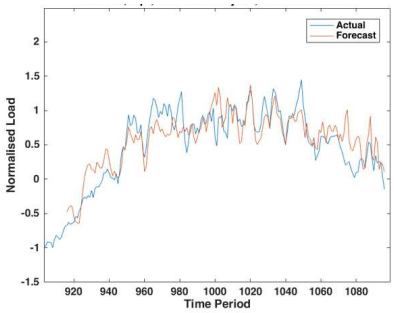

Figure 5. Neural Network regression.

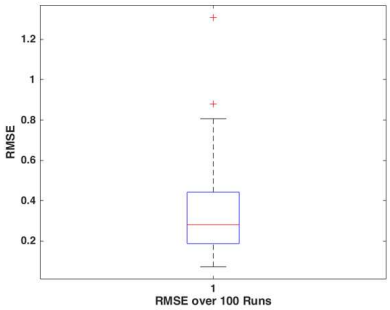

Figure 6. RMSE of Neural Network over 100 runs.

4.3. NARX Gaussian Process

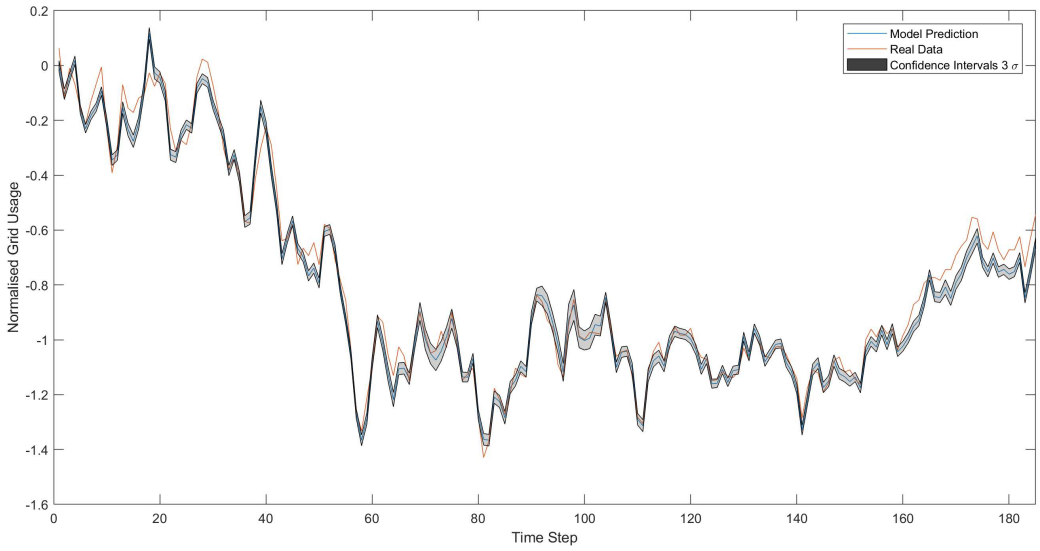

Figure 7. NARX Gaussian Process regression with 3σ confidence intervals.

The best results came from the NARX Gaussian Process which had a prediction error of 0.26%, far outperforming the other algorithms under investigation.

4.4. Evaluation

The linear regression was less effective than the Gaussian process in its forecasting power however it was very quick and simple to run, making it a useful tool for lower accuracy forecasting with a large quantity of data.

The Neural Network for the low number of data points had low prediction accuracy relative to the other methods tested, was slow to run and was difficult to optimise. With more data, it may be more effective especially when compared to a Gaussian process which could run prohibitively slowly on large datasets. NARX Neural Networks could be used and optimised to determine whether their enhancement to be auto-regressive with exogenous variables significantly increases their prediction accuracy.

It was seen in Figure 7 that the Gaussian accurately found the physics of the system, demonstrating the power of auto-regressive gaussian processes for non-stationary regression problems. In an industrial application, the method may be less well suited as it has significant problems of computational time for large numbers of datapoints, being an $\mathcal{O}\left(n^3\right)$ algorithm where n is the number of data points.

References

[1] Soliman S A and Alkandari A M 2010 *Electrical load forecasting: modeling and model construction* (Butterworth-Heinemann) ISBN 0123815444

[2] Stram B N 2016 *Energy Policy* **96** 728–734 ISSN 0301-4215 URL `https://www.sciencedirect.com/science/article/pii/S0301421516302646`

[3] Bishop C M 2006 *Pattern recognition and machine learning* (Springer) ISBN 9780387310732

[4] Rasmussen C E and Williams C K I 2006 *Gaussian processes for machine learning* (MIT Press) ISBN 026218253X URL `https://dl.acm.org/citation.cfm?id=1162254`