# DAILY PEAK ELECTRICITY LOAD FORECASTING IN SOUTH AFRICA USING A MULTIVARIATE NON-PARAMETRIC REGRESSION APPROACH

C. Sigauke[1*], D. Chikobvu[2]

[1*]Department of Statistics and Operations Research, School of Mathematical and Computer Sciences, University of Limpopo, Turfloop Campus, P. Bag X1106, Sovenga, 0727. South Africa
[2]Department of Mathematical Statistics and Actuarial Science, University of the Free State, P. O. Box 339, Bloemfontein 9300, South Africa

Abstract

The paper presents a multivariate adaptive regression splines (MARS) modelling approach for daily peak electricity load forecasting in South Africa for the period 2000 to 2009. MARS is a non-parametric multivariate regression method which is used to solve high dimensional problems with complex model structures such as nonlinearities, interactions and missing data in a straight forward manner and produces results which can be explained to management. The developed model consists of components that represent calendar and meteorological data. The performance of the model is evaluated by comparing it with a piecewise linear regression model. The results from the study show that the MARS model produces better forecast accuracy. Accurate prediction of daily peak load demand is very important for decision makers in the energy sector. This helps in the determination of consistent and reliable supply schedules during peak periods. Accurate short term load forecasts will enable effective load shifting between transmission substations, scheduling of startup times of peak stations, load flow analysis and power system security studies.

Key words: MARS, temperature, peak demand forecasting, piecewise linear regression.

## 1. Introduction

One of the most weather sensitive sectors of any economy is the energy sector. In this sector accurate prediction of daily peak electricity demand is very important. It provides short term forecasts which are required for dispatching and economic grid management of electric energy [1,2,3,8,16,19,21,22]. The most important weather factors which affect daily peak demand (DPD) is temperature. Changing weather conditions represent the major source of variation in peak demand forecasting and the inclusion of temperature has a significant effect due to the fact that in winter, heating systems are used whilst in summer air conditioning appliances are used [6,10,11,13,14,15,17,18]. Other weather factors include: relative humidity, wind speed and cloud cover. Electricity demand forecasting has been studied extensively using various techniques ranging from classical time series methods, neural networks to regression methods. In this paper a multivariate adaptive regression splines model is developed and used to predict daily peak electricity demand for South Africa. An updated review of different forecasting methods can be found in [9,20].

The rest of the paper is organized as follows, in Section 2 the data used is described and a preliminary analysis carried out. The piecewise linear regression and the MARS models are presented in Section 3. A discussion of the results is presented in Section 4 and Section 5 concludes.

## 2. Definitions and Data

The data considered in this paper is on net energy sent out (NESO) from Distribution in response to some demand of electrical power. NESO (measured in megawatts) is defined as the rate at which electrical energy is delivered to customers. In this paper NESO is used as a proxy of electrical demand after adjusting for energy losses. The data is for the period 2000 to 2009.
This definition of electrical demand has its weaknesses. Electrical demand is bounded by the power plants' capacity to provide supply at any time of the day including the need for reserve capacity. Demand cannot exceed supply and there are no market forces acting to influence electricity prices and hence reducing demand in the short run. Prices are generally fixed. If demand were to exceed supply, intervention takes place in the form of, for example, load shedding. Load shedding is the last resort used to prevent a system-wide blackout. This NESO definition excludes the demand from people, companies, etc, who are willing (or unwilling) and able to (or are unable) pay for electricity but currently do not have access to electrical power. Despite the weakness in the NESO definition of electrical demand, it is still a good and measurable proxy for electrical demand.

The daily peak demand (DPD) is the maximum hourly demand in a 24-hour period. Aggregated DPD data was used for the industrial, commercial and domestic sectors of South Africa. The historical data on temperature was collected from 22 meteorological stations from all the provinces of the country. The data was aggregated to get average daily, maximum and minimum temperatures for the whole country.
The time series plot of DPD in Figure 1 shows a positive linear trend and a strong seasonal fluctuation. The trend is mainly due to economic development of the country. Figure 2 shows monthly and daily index plots. The basis for each index is 100. The seasonal peak is in July which is a winter month. There is another small summer peak in October. The daily index plot shows that demand for electricity during the week days is above average consumption and decreases significantly on Saturday and Sunday. A better representation of the relationship between daily peak demand with temperature is shown in Figure 3. Figure 3 shows the relationship between DPD with peak temperature (in degrees Celsius). The peak temperature is the temperature recorded during the hour of peak demand on day $t$. The relationship is nonlinear. The demand of electricity is highly sensitive to temperature fluctuations in winter and less sensitive in summer. DPD increases sharply as temperature decreases. The non-linear relationship between temperature and DPD calls for derivation of two functions, one for the cooling degree-days and the other for heating degree-days. The cooling degree-days ($\text{CDD}_t$) and the heating degree-days ($\text{HDD}_t$) will be estimated on the basis of the following two linear functions as defined in [12],

$\text{CDD}_t = \max(T_t - T_{ref}, 0)$

and

$\text{HDD}_t = \max(T_{ref} - T_t, 0)$

where $T_{ref}$ represents the temperature which separates the winter and summer periods of DPD – temperature relationship and $T_t$ represents the peak temperature on day $t$. The reference temperature ($T_{ref}$) has been selected to be equal to 20.5$^o$C from Figure 3. This appears to be the temperature at which we get minimum demand of electricity. This temperature will be used for the calculation of heating and cooling degree – days. Above this temperature, electricity demand tends to rise slightly and below this temperature electricity demand increases significantly.

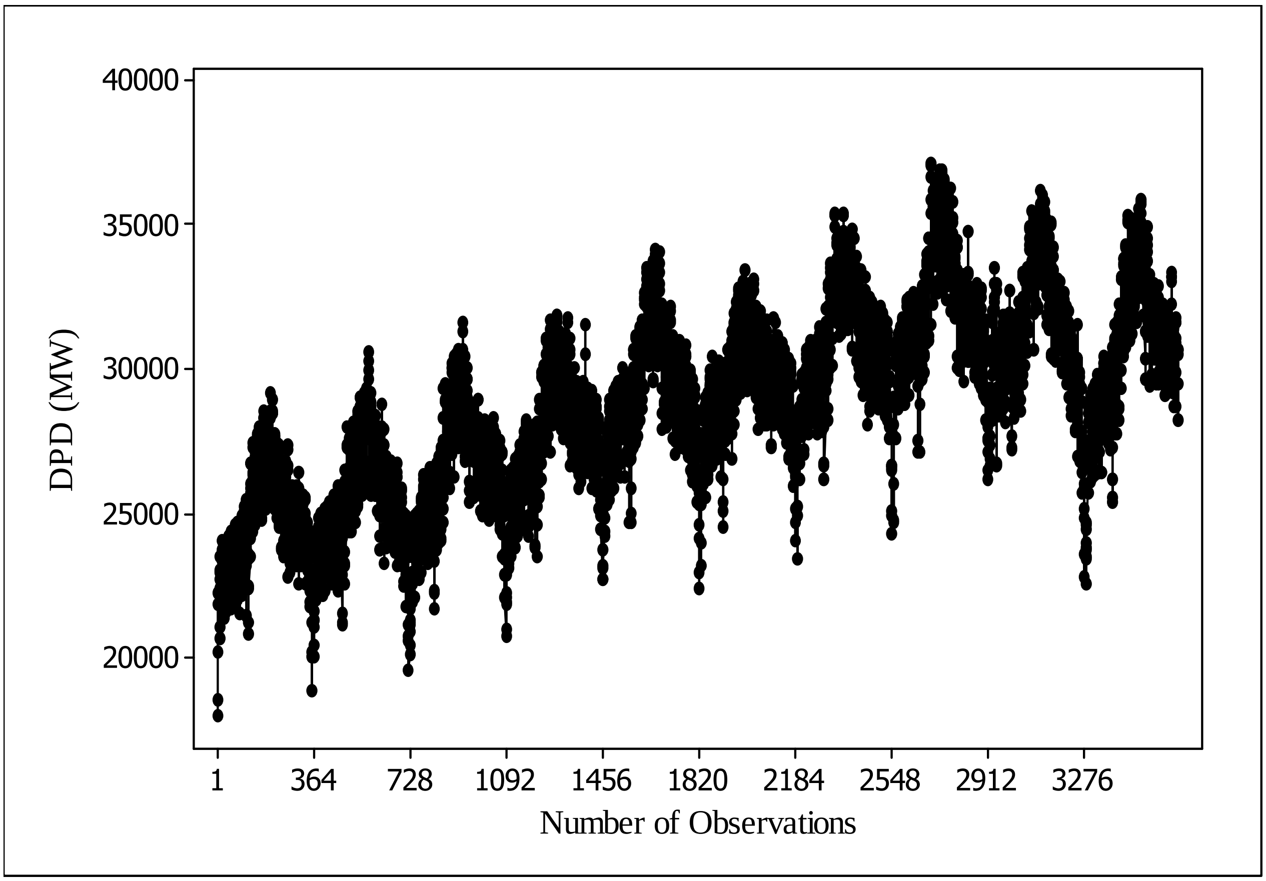
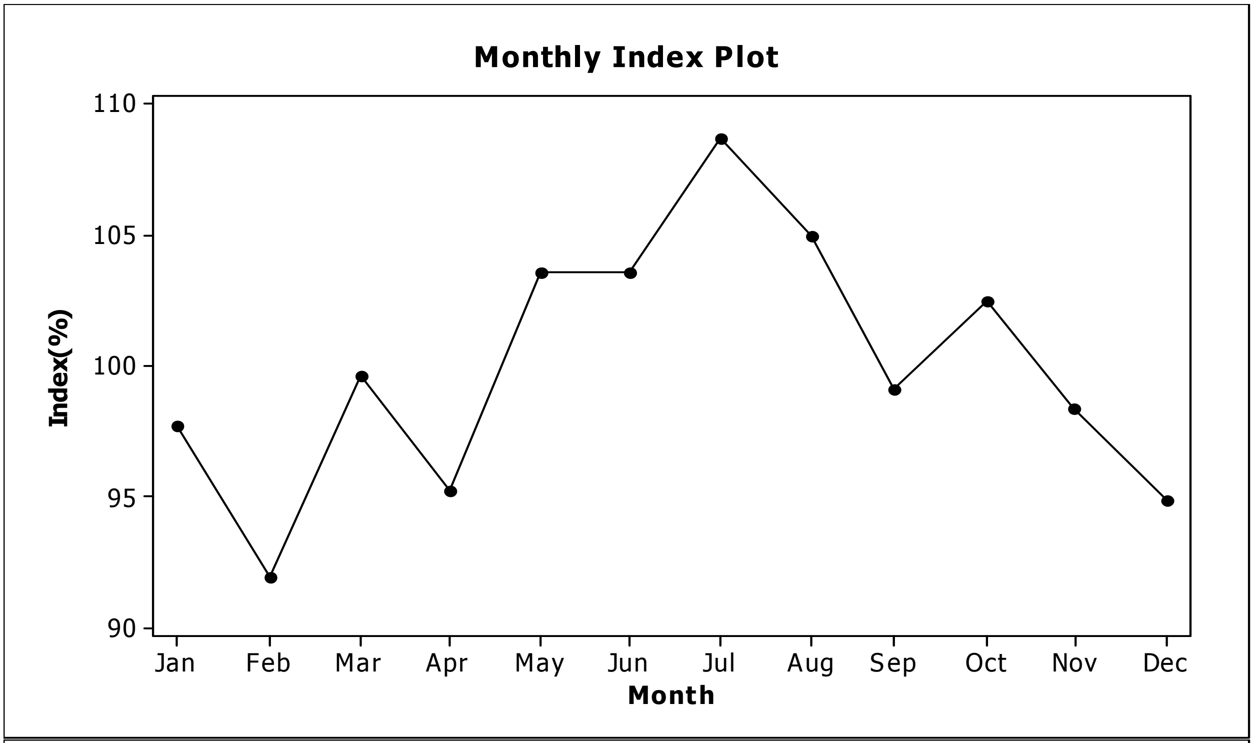



Figure1: Time series plot of daily peak electricity demand for the period 1/1/2000 – 14/12/2009



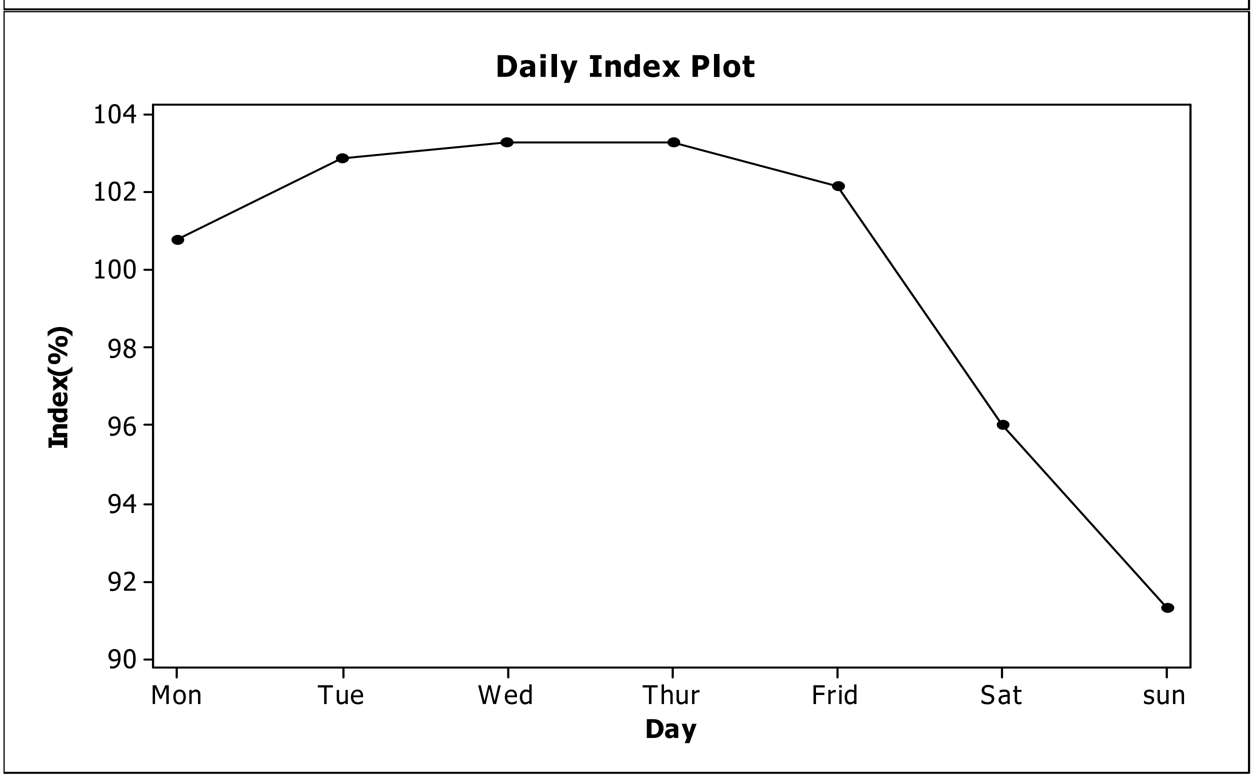




Figure 2: Monthly and daily index plots

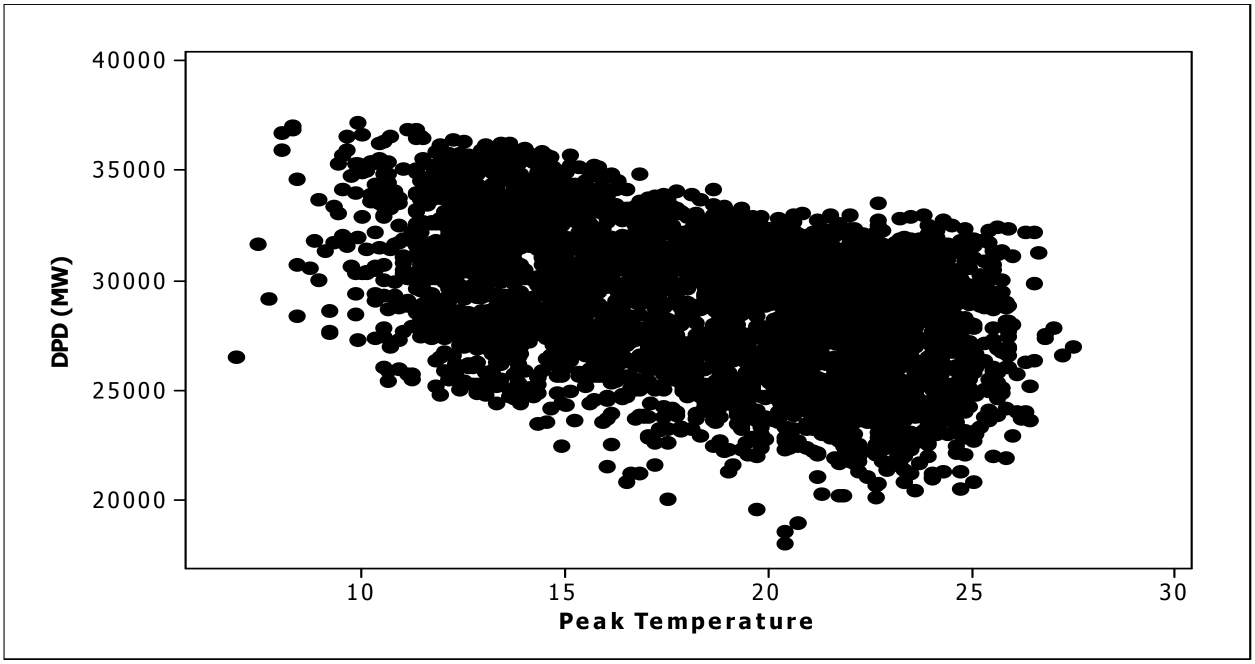



Figure 3: Scatter plot of daily peak demand against peak temperature (in $^{o}$C)

## 3. The Models

A piecewise linear regression model and a multivariate adaptive regression splines model are presented in this section. The developed models are then used for out of sample predictions of DPD. In both models DPD is taken as the dependent variable. The data was transformed by taking natural logarithms to reduce the impact of heteroskedasticity that may be present because of the large data set and its high frequency [13].

### 3.1 The Piecewise Linear Regression Model

Regression based methods have been used extensively in load demand forecasting [4,9,20,22]. They range from simple linear to multivariate linear regression models. These methods work very well when the relationship between the dependent variable and the predictor variables is linear. They are usually fast, reliable and easy to implement with relatively robust solutions.

The relationship between electricity demand and temperature is nonlinear as shown in Figure 3. This calls for use of a multivariate linear regression model with 3 piecewise linear regression functions. These regressions will be representing the winter, non-weather and summer sensitive components.

The piecewise linear regression model used in this paper can be written as

$$z_t = \beta_0 + \beta_1 \mathrm{t} + \beta_2 (x_{pt} - t_w) x_{1t} + \beta_3 (x_{pt} - t_s) x_{2t} + \sum_{d=1}^{7} \alpha_d \mathrm{D}_{dt} + \sum_{j=1}^{12} \tau_j \mathrm{M}_{jt} + \mu \mathrm{H}_t + \delta \mathrm{H}_{t-1} + \lambda \mathrm{H}_{t+1} + R_t \quad (1)$$

where $x_{pt}$ represents peak temperature (in degrees Celsius). The peak temperature is the temperature recorded at the hour of peak demand on day $t$, $z_t$ denotes daily peak demand (in megawatts) observed on day $t$, $t_w$ temperature to identify where the winter sensitive portion of demand join the non-weather sensitive demand component, $t_s$ temperature to identify where the summer sensitive portion of demand join the non-weather sensitive demand component, $\beta_0$ represents the mean daily peak demand observed in the non-weather sensitive period $(t_w \le x_{pt} \le t_s)$. It should be noted that daily peak demand during non-weather sensitive days does not depend on temperature $(x_{pt})$.The variable $t$ represents the trend component, $\mathrm{H}_t$, $\mathrm{H}_{t-1}$ and $\mathrm{H}_{t+1}$ are dummy variables representing holiday, day before and after a holiday respectively. The day of the week effect is represented by $\mathrm{D}_{dt}$, where $d$ represents the days Tuesday up to Sunday with Monday as the base period. $\mathrm{D}_{dt}$ equals 1 if day $d$ is found in observation $t$ and zero otherwise with $t = 1,2,...,n$. The monthly effect is represented by $\mathrm{M}_{jt}$, where $j$ represents the months February up to December with January as the base month. $\mathrm{M}_{jt}$ equals 1 if month $j$ is found in observation $t$ and zero otherwise with $t = 1,2,...,n$,

$R_t = \phi_1 R_{t-1} + \phi_2 R_{t-2} + \phi_5 R_{t-5} + \phi_7 R_{t-7} + \varepsilon_t$, where $R_t$ is a stochastic disturbance term and $\varepsilon_t$ is the innovation in the disturbance with

$$x_{1t} = \begin{cases} 1, & if \quad x_{pt} - t_w < 0 \\ 0, & otherwise \end{cases} \quad \text{and}$$

$$x_{2t} = \begin{cases} 1, & if \quad x_{pt} - t_s > 0 \\ 0, & otherwise \end{cases}$$

Model (1) will account for any residual correlation that may occur as a result of the week to week variation in peak demand and also for the day to day variation. Model (1) is based on the following theoretical assumptions:

1. Peak demand on day $t$ will be highly correlated with peak demand on day $t+1$.
2. There may be significant correlation between demand 2 days, 5 days and/or 7 days apart.

Derivation of the equations of the 3 demand – temperature lines are shown in appendix A1.

## 3.2 Multivariate Adaptive Regression Splines (MARS) Model

Multivariate adaptive regression splines (MARS) is a non-parametric multivariate regression method which was developed in [8]. MARS has been used to solve high dimensional problems with complex model structures such as nonlinearities, interactions, multicollinearity and missing values [3, 11, 17, 27]. The method does not make any assumptions about the functional relationship between the response variable and the predictor variables. The MARS modeling approach overcomes the major drawbacks of using artificial neural networks which are long training processes, interpretive difficulties and the inability to determine the relative importance of potential input variables. The MARS algorithm divides the modeling space into subregions and then fits in each subregion simple linear regression models. The model building process is in two steps, the forward stepwise algorithm and the backward stepwise algorithm. In the forward stepwise step the MARS algorithm constructs a large number of basis functions which will over fit the data. In the backward stepwise algorithm basis functions are deleted in order of least contribution using the generalized cross validation (GCV) criterion. The general MARS model can be written as in [8],

$$f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m B_m(x) \qquad (2)$$

where $B_m(x)$ is a basis function written as

$$B_m(x) = \prod_{k=1}^{K_m} \left[ s_{km} \left( x_{v(k,m)} - t_{km} \right) \right],$$

$\alpha_0$ and $\alpha_m$ are parameters, $M$ is the number of basis functions, $K_m$ is the number of knots, $s_{km}$ takes on values of either 1 or -1 and indicates the right or left sense of the associated step function, $v(k,m)$ is the label of the independent variable and $t_{km}$ indicates the knot location.

The MARS algorithm will then select variables and values of those variables for knots of the hinge functions.

In order to choose the best subset model using MARS a generalized cross validation (GCV) criterion is used. It is a measure of the goodness of fit which takes into account the residual error and the model complexity. In its simplest form the generalized cross validation criterion can be written as in [5],

$$\text{GCV}(M) = \frac{1}{N} \sum_{i=1}^{N} \left[ y_i - \hat{f}_M(x_i) \right]^2 \Big/ \left[ 1 - \frac{C(M)}{N} \right]^2 \qquad (3)$$

where $N$ is the sample size, $C(M)$ is the cost-penalty measures of a model containing $M$ basis functions. The numerator measures the lack of fit on the $M$ basis function model $\hat{f}_M(x_i)$ and the denominator represents the penalty for the model complexity $C(M)$. The complexity cost function can be written as in [8],

$$C(M) = \text{trace}(\text{B}(\text{B}^T\text{B})^{-1}\text{B}^T) + 1,$$

where $\text{B}$ is the $M \times N$ data matrix of the $M$ (nonconstant) basis functions $(B_{ij} = B_i(x_j))$. The best model is one with the lowest GCV.

The three general MARS models used in this paper are shown in equations (4), (5) and (6) respectively.

$$z_t = a_0 + c_1 \max(0, x_{pt} - t_w) + c_2 \max(0, x_{pt} - t_s) + \sum_{d=1}^{7} \alpha_d \text{D}_{dt} + \sum_{j=1}^{12} \tau_j \text{M}_{jt} + \mu \text{H}_t + \delta \text{H}_{t-1} + \lambda \text{H}_{t+1} + \varepsilon_t \qquad (4)$$

$$z_t = \omega_0 + c_3 \max(0, T_t - T_{ref}) + c_4 \max(0, T_{ref} - T_t) + \sum_{d=1}^{7} \alpha_d \text{D}_{dt} + \sum_{j=1}^{12} \tau_j \text{M}_{jt} + \mu \text{H}_t + \delta \text{H}_{t-1} + \lambda \text{H}_{t+1} + \varepsilon_t \qquad (5)$$

$$z_t = \beta_0 + c_5 \max(0, T_t - T_{ref}) + c_6 \max(0, T_{ref} - T_t) + \varepsilon_t \qquad (6)$$

where $a_0$, $\omega_0$, $\beta_0$ and $c_1,...,c_6$ are constants, $z_t$ represents daily peak demand, $x_{pt}$ represents peak temperature (in degrees Celsius), $t_w$ temperature to identify where the winter sensitive portion of demand join the non-weather sensitive demand component; $t_s$ temperature to identify where the summer sensitive portion of demand join the non-weather sensitive demand component, $T_{ref}$ represents the temperature which separates the winter and summer periods of DPD – temperature relationship and $T_t$ represents the daily average temperature for day $t$.

## 4. Results and Discussion

### 4.1 Piecewise Linear Regression Model

Three different piecewise linear functions for modeling the peak demand $(z_t)$ and peak temperature $(x_{pt})$ relationship were proposed in model (1). The values of $t_w$ and $t_s$ are 17.5°C and 24°C respectively. These values were determined from visual inspection of the graph in Figure 3. Piecewise linear regression models were run for various reference temperatures in the interval 17 °C - 24°C, without any significant improvements in the results. The reference temperature ($T_{ref}$) has been selected to be equal to 20.5°C from Figure 3. The piecewise linear function is shown in equation (7).

$$z_t = \beta_0 + \beta_1 t + \beta_2(x_{pt} - 17.5)x_{1t} + \beta_3(x_{pt} - 24)x_{2t} + \sum_{d=1}^{7}\alpha_d \mathrm{D}_{dt} + \sum_{j=1}^{12}\tau_j \mathrm{M}_{jt} + \mu \mathrm{H}_t + \delta \mathrm{H}_{t-1} + \lambda \mathrm{H}_{t+1} + \mathrm{R}_t$$

where $R_t = \phi_1 R_{t-1} + \phi_2 R_{t-2} + \phi_5 R_{t-5} + \phi_7 R_{t-7} + \varepsilon_t$. (7)

Out of the 3636 data points, which is from 1 January 2000 to 14 December 2009, 3592 data points (1 January 2000 to 31 October 2009) were used for developing the models and the remaining 44 observations were then used for validation. Table 1 shows the parameter estimates of the best piecewise linear regression model. The model can be written as

$$\begin{aligned} z_t = {} & 26274 + 1.9\mathrm{t} - 232.8(x_{pt} - 17.5)x_{1t} + 21.0(x_{pt} - 24)x_{2t} - 881.7\mathrm{D}_{5t} - 2279.0\mathrm{D}_{6t} \\ & - 2551.6\mathrm{D}_{7t} - 1813.4\mathrm{H}_t - 810.9\mathrm{H}_{t-1} - 248.7\mathrm{H}_{t+1} + 0.819 z_{t-1} + 0.013 z_{t-2} + 0.056 z_{t-5} + 0.094 z_{t-7} \end{aligned} \tag{8}$$

The coefficient of $t$ is positive showing a positive linear trend. The dummy variable $x_{1t}$ which was defined as

$$x_{1t} = \begin{cases} 1, & if \quad x_{pt} - t_w < 0 \\ 0, & otherwise \end{cases}$$

in equation (1) is positive showing that if temperature decreases by one degree from 17.5°C, electricity demand will increase by 232.8 MW. The coefficient of $x_{2t}$ which was also defined as

$$x_{2t} = \begin{cases} 1, & if \quad x_{pt} - t_s > 0 \\ 0, & otherwise \end{cases}$$

is also positive showing that if temperature increases by one degree from 24°C, electricity demand will increase by 21 MW. This shows that electricity demand is more sensitive to winter than summer. All the coefficients of the dummy variables representing Friday, Saturday, Sunday, holiday, day before holiday and day after holiday are negative. This shows that there is a decrease in demand during these periods. The largest decrease is on Sunday out of the three days of the week. During holidays, demand for electricity decreases significantly compared to a day before and after a holiday. The least decrease is experienced a day after a holiday.

Table 1: Parameter estimates of the piecewise linear regression model

| Par | $c$ | $H_t$ | $H_{t-1}$ | $H_{t+1}$ | Friday | Saturday | Sunday |
|---|---|---|---|---|---|---|---|
| Coef | 26274.0 (0.000) | -1813.4 (0.000) | -810.9 (0.000) | -248.7 (0.000) | -881.7 (0.00) | -2279.0 (0.000) | -2551.6 (0.000) |
| Par | $t$ | $x_{1t}$ | $x_{2t}$ | $\phi_1$ | $\phi_2$ | $\phi_5$ | $\phi_7$ |
| Coef | 1.9 (0.000) | 232.8 (0.000) | 21.0 (0.4963) | 0.819 (0.000) | 0.013 (0.4679) | 0.056 (0.000) | 0.094 (0.000) |

## 4.1 The MARS Models

### 4.1.1 Model 1

Model (1) is a simple MARS model which was used to determine the reference temperature which seperates the winter from the summer periods of the daily peak demand – temperature relationship. DPD was the dependent variable with the peak temperature as the regressor variable. The best MARS model had a GCV value of $8.66744\times10^{6}$ and the reference temperature was found to be 20.9$^{o}$C. This is given in equations (9) and (10) respectively.

$$z_t = \omega_0 + c_3\max(0, x_{pt} - T_{ref}) + c_4\max(0, T_{ref} - x_{pt}) \tag{9}$$

The basis functions are
BF1 = max (0, Xpt - 20.9);
BF2 = max (0, 20.9 - Xpt);
Z = 27833.6 - 125.423 * BF1 + 384.209 * BF2;

and the complete model can be written as

$$\text{DPD} = 27833.6 - 125.423\max(0, x_{pt} - 20.9) + 384.209\max(0, 20.9 - x_{pt})\ . \tag{10}$$

If temperature decreases by a degree from 20.9$^{o}$C, DPD will increase by 384.209MW. Similarly an increase by one degree above 20.9$^{o}$C will result in a decrease of 125.423MW. This shows that DPD is more sensitive to low temperatures. This model is used to determine the number of heating degree days and also the number of cooling degree days. It cannot be used for predictions since only one predictor variable was used which is peak temperature.

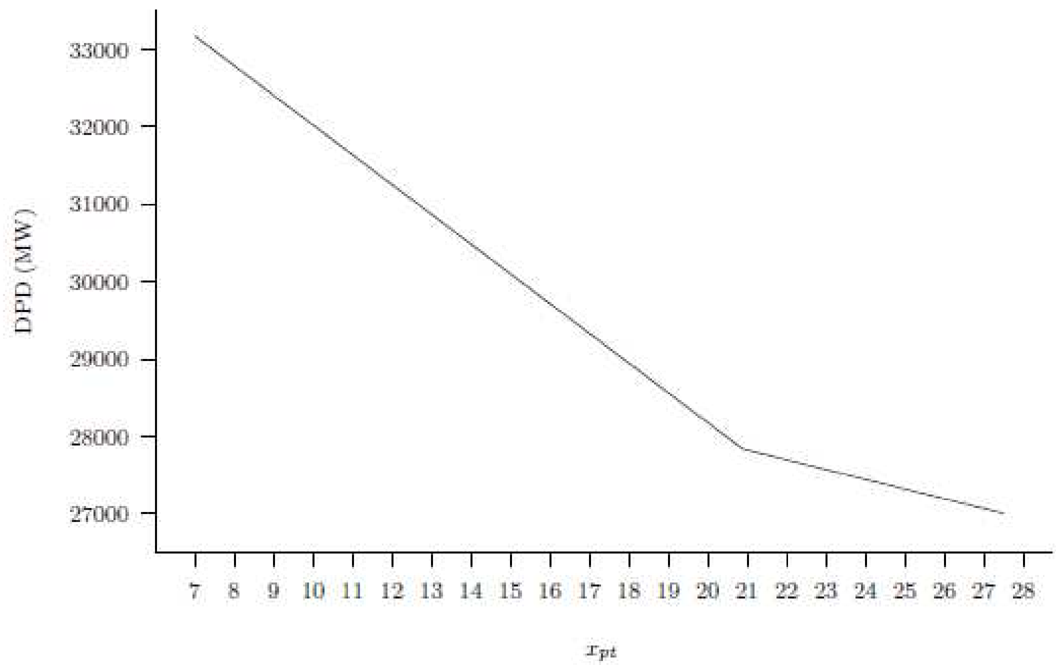



Figure 4: MARS plot of model 1

#### 4.1.2 Model 2

Out of the 24 predictor variables the MARS algorithm selected 8 as the most important. These are shown in table 2 in order of their importance. The piecewise linear GCV was $9.02477 \times 10^5$.

Table 2: Important predictor variables using MARS

| Variable | Importance | GCV |
|---|---|---|
| t | 100.000 | 7.27027E+06 |
| $x_{pt}$ | 58.25802 | 3.06371E+06 |
| $D_{7t}$ | 34.58035 | 1.66394E+06 |
| $D_{6t}$ | 31.78196 | 1.54569E+06 |
| $H_t$ | 18.85743 | 1.12892E+06 |
| $H_{t-1}$ | 14.80247 | 1.04201E+06 |
| $D_{5t}$ | 12.35992 | 9.99758E+05 |
| $H_{t+1}$ | 8.76643 | 9.51416E+05 |

The basis functions are

BF1 = max (0, t - 2735);
BF2 = max(0, 2735 - t );
BF3 = max(0, $x_{pt}$ - 19.2);
BF4 = max(0,19.2 - $x_{pt}$ );
BF5 = ( $D_{7t}$ = 0);
BF7 = ( $D_{6t}$ = 1);
BF9 = ( $H_t$ = 0);
BF11 = ( $D_{5t}$ = 0);
BF13 = ( $H_{t-1}$ = 0);
BF15 = ( $H_{t+1}$ = 0) * BF13;
Z = 24676.4 - 1.74202 * BF1 - 2.9918 * BF2 - 235.381 * BF3 + 411 * BF4 + 2598.11 * BF5 - 2367.12 * BF7 + 2722.32 * BF9 + 924.862 * BF11+ 500.434 * BF13 + 1280.82 * BF15;

The final model is given in equation (11).

$$\begin{aligned} \mathrm{DPD} = {} & 24676.4 - 1.74\max(0, t - 2735) - 2.99\max(0, 2735 - t) - 235.38\max(0, x_{pt} - 19.2) \\ & + 411\max(0, 19.2 - x_{pt}) + 2598.11(\mathrm{D_{7t}} = 0) - 2367.12(\mathrm{D_{6t}} = 1) + 2722.32(\mathrm{H_t} = 0) \\ & + 924.86(\mathrm{D_{5t}} = 0) + 500.43(\mathrm{H_{t-1}} = 0) + 1280.82(\mathrm{H_{t+1}} = 0)(\mathrm{H_{t-1}} = 0) \end{aligned} \tag{11}$$

The coefficient of basis function 1 is negative meaning that if trend is above 2375 electricity demand will decrease by 1.74 MW and when its below this knot it will decrease by 2.99MW. Basis function three's coefficient is negative implying that if peak temperature is above 19.2$^0$C DPD will decrease by 235.381MW and if peak temperature is below this knot DPD will increase by 411MW. Coefficient for basis function 5 is positive meaning if the day of the week is not Sunday, DPD will increase by 2598.11MW, but if the day is Saturday there will be a decrease in DPD of 2367.12MW. The coefficient of basis function 9 is positive meaning that if the day $t$ is not a holiday the DPD will increase by 2722.32MW and if it is not a day before a holiday DPD will increase by 500.434MW. There is one bivariate interaction between a day before and a day after a holiday. If day $t$ is not a day before or after a holiday the DPD will increase by 1280.82MW. If day $t$ is not a Friday there will be an increase in DPD of 924.862MW.

### 4.1.3 Model 3

The third model is a MARS model for Average Daily Energy Sent Out (ADESO) with average daily temperature (ADT) as the predictor variable. The model identifies the winter sensitive, weather neutral and summer sensitive periods. The basis functions are

BF2 = max(0, 22 - ADT);
BF3 = max(0, ADT - 16);

Z = 564863 + 7332.94 * BF2 + 3714.8 * BF3;

$$\text{ADESO} = 564863 + 7332.94\max(0, 22 - \text{ADT}) + 3714.8\max(0, \text{ADT} - 16) \qquad (12)$$

The piecewise linear GCV was $3.82422 \times 10^9$. The graphical plot of ADESO against average daily temperature is shown in Figure 4. The three demand-temperature equations are shown in equations 13 – 15.
If average daily temperature is less than or equal to $16^0$C we use

$$\text{ADESO} = 564863 + 7333\max(0, 22 - \text{ADT}) \qquad (13)$$

That is, if the temperature decreases by $1^0$C (e.g. from $16^0$C to $15^0$C) electricity demand will increase by 7333MW, which is about 1.2% increase.

If average daily temperature is greater than or equal to $22^0$C we use

$$\text{ADESO} = 564863 + 3715\max(0, \text{ADT} - 16) \qquad (14)$$

If temperature increases by $1^0$C (e.g. from $22^0$C to $23^0$C) electricity demand will increase by 3715MW, which is about 0.6% increase.

For the average daily temperature between $16^0$C - $22^0$C we use

$$\text{ADESO} = 564863 + 7333\max(0, 22 - \text{ADT}) + 3715\max(0, \text{ADT} - 16) \qquad (15)$$

If temperature decreases by $1^0$C in the range $16^0$C - $22^0$C (e.g. from $22^0$C to $21^0$C) electricity demand will increase by 3618MW, which is about 0.6% increase.
The graphical plot for average daily energy sent out against average daily temperature and the MARS plot are shown in Figure 5 and 6 respectively. The modeling space is divided into three subspaces separated by two knots at $16^0$C and $22^0$C as shown in Figure 6.

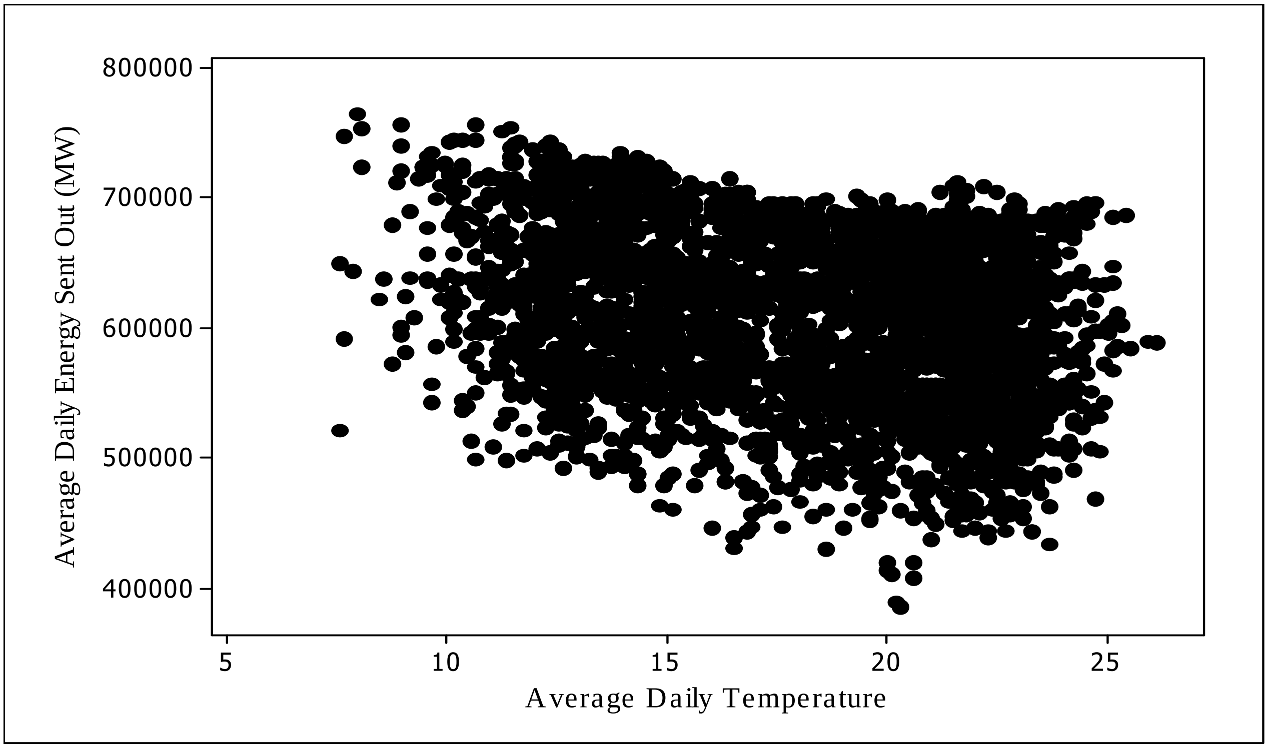



Figure 5: Scatter plot of average daily energy sent out against average daily temperature

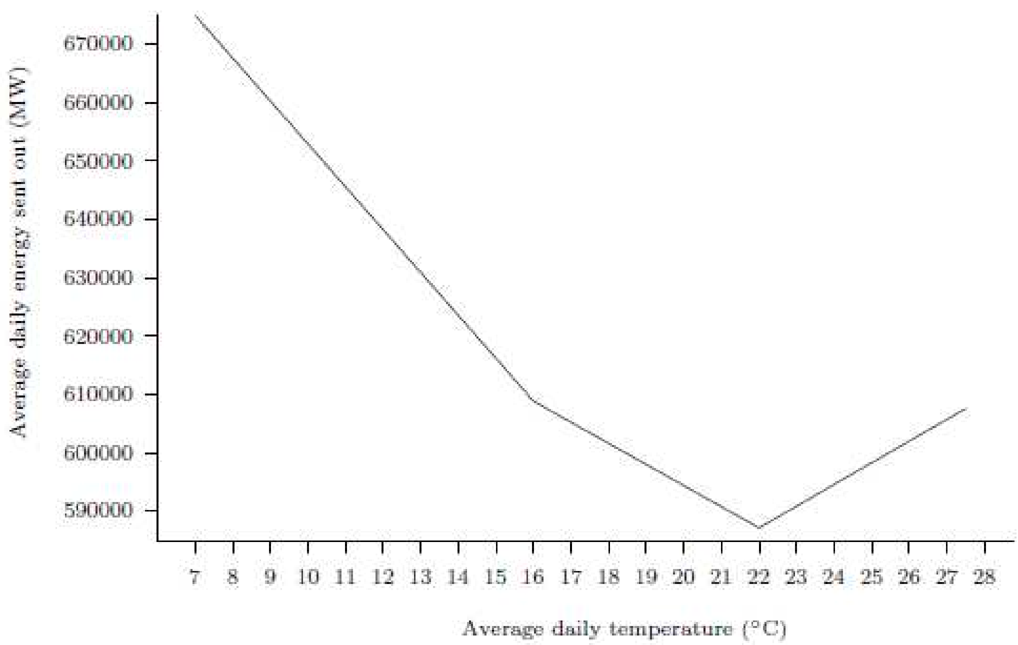



Figure 6: MARS plot of model 3

4.2 Evaluating the goodness of fit of the models

The root mean squared error (RMSE) is used for the evaluation of the piecewise regression model and the MARS model for peak load demand forecasting in the out of sample predictions for the period 1 November to 14 December 2009. The training period was 1 January 2000 to 31 October 2009. The RMSE is calculated as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}\left(z_{at} - z_{ft}\right)^2}{n}} \tag{16}$$

where $n$ is the number of out of sample forecast data points and $z_{at} - z_{ft}$ represents the forecast errors. The terms $z_{at}$ and $z_{ft}$ are the actual DPD and its future forecast respectively. Table 3 presents the comparative performances of the piecewise and MARS model in terms of testing error. The MARS model outperformed the piecewise linear regression model.

Table 3: Comparative Analysis results of the piecewise and MARS models

| | R-Squared Adjusted | RMSE<br>Validation (Testing Period) |
|---|---|---|
| Piecewise Linear | 0.91626 | 940.84331 |
| MARS Model 2 | 0.980964 | 446.013 |

## 5. Conclusions

A MARS model was developed for predicting daily electricity peak demand and the performance of the model was compared with a piecewise linear regression model. There were 3636 data points, from 1 January 2000 to 14 December 2009. 3592 data points were used for developing the models and the remaining 44 observations were then used for validation. The MARS model outperformed the piecewise linear regression model. The developed MARS model is easy to explain to management. The model is capable of clustering together categories of variables that have similar effects on the dependent variable. Accurate prediction of daily peak load demand is very important for decision makers in the electricity sector. This helps in the determination of consistent and reliable supply schedules during peak periods. Accurate short term load forecasts will enable effective load shifting between transmission substations, scheduling of startup times of peak stations, load flow analysis and power system security studies. Future research should look at the investigation of weather sensitivity analysis on daily and seasonal peak electricity demand done on each of the provinces of South Africa. Another interesting area to investigate would be the development of a hybrid model which will integrate the multivariate adaptive regression splines approach with neural network techniques and also the use of other adaptive techniques such as classification and regression trees (CART), TreeNet and Random Forests. These studies will be done elsewhere.

## Acknowledgments

The authors are grateful to Eskom for providing the data, Department of Statistics and Operations Research, School of Mathematical and Computer Sciences, University of Limpopo, Turfloop Campus and Department of Mathematical Statistics and Actuarial Science, University of the Free State for using their resources and to the numerous people who assisted in making comments on this paper.

## References

[1] Amaral LF, Souza RC & Stevenson M, 2008, *A smooth transition periodic autoregressive (STPAR) model for short-term load forecasting,* International Journal of Forecasting, **24**, pp. 603-615.

[2] Amin-Naseri MR & Soroush AR, 2008, *Combined use of unsupervised and supervised learning for daily peak load forecasting,* Energy Conversion and Management, **49**, pp.1302-1308.

[3] Chen L-A, 1997, *Multivariate regression splines,* Computational Statistics and Data Analysis **26**, pp.71-82.

[4] Chou S-M, Lee T-S, Shao YE & Chen I-F, 2004, *Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines.* Expert Systems with Applications, **27**, pp.133-142.

[5] Craven P & Wahba G, 1979, *Smoothing noisy data with spline functions. Estimating the correctdegree of smoothing by the method of generalized cross-validation,* Numberische Mathematik, 31, pp.317-403.

[6] Feinberg E & Genethliou D, 2005, Load *Forecasting. In J. Chow, F. Wu and J. Momoh (Eds.). Applied Mathematics for Restructured Electric Power Systems* Optimization, Control and Computational Intelligence, pp. 269-285, Springer.

[7] Franco G & Sanstad AH, 2008, *Climate change and electricity demand in Carlifornia. Climatic Change (Supplement 1),* Springer, **87**, pp. S139-S151.

[8] Friedman JH, 1991, *Multivariate adaptive regression splines*, The Annals of Statistics, **19(1)**, pp. 1-141.

[9] Ghosh S, 2008, *Univariate time-series forecasting of monthly peak demand of electricity in northern India*, International Journal of Indian Culture and Business Management, **1**, pp. 466-474.

[10] Goia A, May C & Fusai G, 2010, *Functional clustering and linear regression for peak load forecasting*, International Journal of Forecasting, Article in press.

[11] Gooijer JG De & Ray BK, 2003, Modeling vector nonlinear time series using POLYMARS, Computational Statistics & Data Analysis, **42**, pp. 73-90.

[12] Hahn H, Meyer-Nieberg S & Pickl S, 2009, *Electric load forecasting methods: Tools for decision making*, European Journal of Operational Research, **199**, pp. 902-907.

[13] Hekkenberg M, Benders RMJ, Moll HC & Schoot AJM, 2009, *Indications for a changing electricity demand pattern: The temperature dependence of electricity demand in the Netherland*, Energy Policy, **37**, pp. 1542-1551.

[14] Hekkenberg M, Moll HC & Schoot AJM, Dynamic *temperature dependence patterns in future energy demand models in the context of climate change*. Energy, **34**, pp. 1797-1806.

[15] Makridakis S, Wheelwright S & Hyndman, 1998, *Forecasting methods and applications*, 3rd Edition, Wiley, New York.

[16] Mirasgedis S, Sarafidis Y, Georgopoulou E, Lalas DP, Mschovitis M, Karagiannis F & Papakonstantinou D, 2006, *Models for mid-term electricity demand forecasting incorporating weather influences*, Energy, **31**, pp. 208-227.

[17] Munoz J & Felicisimo AM, 2004, *Comparison of statistical methods commonly used in predictive modeling*, Journal of Vegetation Science, **15**, pp. 285-292.

[18] Pilli-Sihvola K, Aatola P, Ollikainen M & Tuomenvirta H, 2010, *Climate change and electricity consumption – Witnessing increasing or decreasing use and costs*, Energy Policy, **38(5)**, pp. 2409-2419.

[19] Psiloglou BE, Giannakopoulos C, Majithia S & Petrakis M, 2009, *Factors affecting electricity demand in Athens, Greece and London, UK: A comparative assessment*, Energy, **34**, pp. 1855-1863.

[20] Ismail Z, Jamaluddin F & Jamaluddin F, 2008, *Time series regression model for forecasting Malaysian electricity load demand*, Asian Journal of Mathematics and Statistics, **1(3)**, pp. 139-149.

[21] Ismail Z, Yahya A & Mahpol KA, 2009, *Forecasting peak load electricity demand using statistics and rule based approach*, American Journal of Applied Sciences, **6(8)**, pp. 1618-625.

[22] Ramanathan R, Engle R, Granger CWJ, Vahid-Araghi F & Brace C, 1997, *Short-run forecasts of electricity loads and peaks*, International Journal of Forecasting, **13**, pp. 161-174.

[23] Soares LJ & Medeiros MC, 2008, *Modelling and Forecasting short-term electricity load: A comparison of methods with an application to Brazilian data*, International Journal of Forecasting, **24**, pp. 630-644.

[24] Taylor J, de Menezes L & Mcsharry P, 2006, *A comparison of univariate methods for forecasting electricity demand up to a day ahead.* International Journal of Forecasting, **22**, pp. 1-16.

[25] Taylor JW, 2008, An *evaluation of methods for very short-term load forecasting using minute-by-minute British data*, International Journal of Forecasting, **24**, pp. 645-658.

[26] Truong N-V, Wang L & Wong PKC, 2008, *Modelling and short-term forecasting of daily peak power demand in Victoria using two-dimensional wavelet based SDP models*, Electrical Power and Energy Systems, **30**, pp. 511-518.

[27] Tsekouras GJ, Dialynas EN, Hatziargyriou ND & Kavatza S, 2007, *A non-linear multivariable regression model for midterm energy forecasting of power systems.* Electric Power Systems Research, **77**, pp. 1560-1568.

APPENDIX A1

Derivation of the equations of the 3 demand – temperature lines

$$z_t = \beta_0 + \beta_1 \mathrm{t} + \beta_2 (x_{pt} - t_w) x_{1t} + \beta_3 (x_{pt} - t_s) x_{2t} + \sum_{d=1}^{7} \alpha_d \mathrm{D}_{dt} + \sum_{m=1}^{12} \tau_m \mathrm{M}_{mt} + \mu \mathrm{H}_{jt} + \delta \mathrm{H}_{jt-1} + \lambda \mathrm{H}_{jt+1} + R_t \tag{A1}$$

Let $\sum_{d=1}^{7} \alpha_d \mathrm{D}_{dt} = \sum_{m=1}^{12} \tau_m \mathrm{M}_{mt} = \mu \mathrm{H}_{jt} = \delta \mathrm{H}_{jt-1} = \lambda \mathrm{H}_{jt+1} = 0$ (A2)

Substituting (A2) into (A1) we get three equations for winter-sensitive, summer-sensitive and non-weather sensitive months. The equations are (A3), (A4) and (A5) respectively.

Winter – sensitive months

ie $x_{pt} < t_w, \quad x_{1t} = 1, \quad x_{2t} = 0$ we get

$$E(z_t) = \beta_0 + \beta_1 (x_{pt} - t_w)(1) + \beta_2 (x_{pt} - t_s)(0)$$

$$E(z_t) = \beta_0 + \beta_1 (x_{pt} - t_w) \tag{A3}$$

$$E(z_t) = (\beta_0 - \beta_1 t_w) + \beta_1 x_{pt}$$

Summer – sensitive months

ie $x_{pt} > t_s, \quad x_{1t} = 0, \quad x_{2t} = 1$ we get

$$E(z_t) = \beta_0 + \beta_1 (x_{pt} - t_w)(0) + \beta_2 (x_{pt} - t_s)(1)$$

$$E(z_t) = \beta_0 + \beta_2 (x_{pt} - t_s) \tag{A4}$$

$$E(z_t) = (\beta_0 - \beta_2 t_s) + \beta_2 x_{pt}$$

Non – weather sensitive months

ie $t_w \le x_{pt} \le t_s, \quad x_{1t} = x_{2t} = 0$ we get

$$E(z_t) = \beta_0 \tag{A5}$$

Where $\beta_0$ represents the mean daily peak demand observed in the non-weather sensitive period.