Improvement of electricity consumption forecasts using temperature inputs

Fiona T. Murray *, John V. Ringwood

School of Electronic Engineering, Dublin City University, Dublin 9, Ireland

Received 25 March 1994; revised 2 September 1994

Abstract

Forecasting electricity consumption requires the projection of past consumption patterns into the future using a mathematical model. The simplest model for doing this is the univariate model, which has no inputs, and therefore requires no knowledge of future trends in variables affecting consumption. It is well known that input driven (causal) models can have significantly better performance than univariate models, particularly when the input sequences are known. The paper demonstrates the improvement obtained in the case of a temperature input. Two case studies are examined, these case studies are carefully selected, since they represent situations where the temperature variable is a dominant and non-dominant input. It is shown that increased accuracy in the forecasting results can be achieved even when a non-dominant input is used. The analysis includes the determination of the optimal configuration of the temperature input and this is justified for two case studies examined. It is also shown that when future inputs are unknown, using a separate model to estimate future values of the input (and subsequent use of these estimates as inputs to the forecasting model) results in improved electricity consumption forecasts. In addition, two methods for the determination of the electricity forecasting models are employed.

Key words: Causal forecasting model; Electricity demand; Correlation analysis; Dominant and non-dominant inputs

1. Introduction

This paper deals with the time series modelling approach to forecasting weekly electricity consumption. This method involves the analysis of past data, which facilitates the construction of mathematical models which can accurately predict future electricity demand. The simplest mathematical models, univariate models, utilise past electricity consumption data; they do not take into account the affect of
other variables which influence electricity demand. Therefore the use of these models require no knowledge of future trends in variables affecting consumption. However the paper shows that even when future inputs are unknown and it is necessary to predict these values, increased accuracy can be achieved by including the affect of these exogenous inputs in the mathematical model. The paper demonstrates the improvement obtained in the case of using a temperature input since this tends to be a very significant input for the time scale under study, i.e. forecasting weekly electricity demand one year in advance. The study is carried out on data obtained from two notably different power boards operating in diverse global areas; hence it is possible to form a generalised conclusion. The first of these electricity utilities is the Irish national power board operating an isolated network with peak load of 2500 MW. The second utility is a regional power board in Northern New Zealand, operating a peak load of 14 MW. The Irish and New Zealand climates are described broadly by the same climate region classification, i.e. Temperate Oceanic, although due to the closer proximity of New Zealand to the equator the temperature here is moderately higher. Weather affects electricity consumption in each system primarily through heating requirements, cooling requirements (air conditioning etc.) are not a major factor. In the specific region of New Zealand in question, weather, although a major influencing factor, is not the dominant factor, electricity demand here is driven primarily by the agricultural and forestry industry. In Ireland on a national scale weather is the dominant factor. In both case studies modelling is performed on total consumption, the data being available on a weekly basis for approximately the past eleven years. The main differences between the two power boards are summarised in Table 1.

The paper presents univariate and causal electricity consumption forecasting models and analyses the results from each model. A comparison of the results demonstrates the improvement in accuracy which can be achieved through the use of appropriate causal models. The causal models are simulated using actual and predicted temperature inputs. Univariate models are built to forecast the temperature input required for the causal model. For each data set, two different modelling techniques are employed to build the models. The first approach uses Box–Jenkins forecasting techniques, a SARIMA (seasonal autoregressive integrated moving average) model is built for the univariate system, and a Box–Jenkins transfer function model is built for the causal system. The second approach constructs an ARMA (autoregressive moving average) model for the univariate system, and an ARMAX (autoregressive moving average with eXogenous inputs) model for the causal system.

<table>
<thead>
<tr>
<th>Northern New Zealand Power Board</th>
<th>Irish Power Board</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regional Power Board</td>
<td>National Power Board</td>
</tr>
<tr>
<td>Peak load 14 MW</td>
<td>Peak load 2000 MW</td>
</tr>
<tr>
<td>Operating in Temperate Oceanic climate</td>
<td>Operating in Temperate Oceanic climate</td>
</tr>
<tr>
<td>Temperature non-dominant input</td>
<td>Temperature dominant input</td>
</tr>
</tbody>
</table>
There are several studies devoted to the use of exogenous variables in forecasting electricity demand. Train [18], Engle [5], Schneider [15], Baker [2] and Gupta [6] all utilise weather variables in the models which they build to forecast electricity demand. Train and Engle both use heating and cooling degree-day (HDD and CDD) variables in their regression models. They in fact include HDD and CDD calculated at different base temperatures in the same model, for the case studies examined in this paper only HDD calculated at one base temperature is included in the causal models; Train and Engle, are however building forecasting models using data from an American power utility where both heating and cooling requirements are significant. Train concentrates on finding the best method of calculating the weather variables to associate with monthly sales billing data and highlights the importance of finding an appropriate form of an input variable from the data available. Schneider when forecasting 24-hour loads utilises a temperature deviation variable which reflects the change in the load due to the weather being different than expected, it is calculated as the expected minus the actual (or forecast) temperature. He also observed the effect that during the heating season there is a base temperature above which a change in temperature had a less significant effect on the load than when the temperature falls below this base temperature. A similar but opposite effect existed during the cooling season. An effective temperature deviation was thus calculated taking this into account; this same effect is taken into account in this paper through the use of the HDD figures. Baker when forecasting daily load uses average temperature, effective temperature, cooling power of the wind and effective illumination of the sky weather variables when fitting regression models to the daily load data. He also accounts for the effect of the existence of a base temperature when modelling the weather/load response. The literature discussed above deals mainly with peak, hourly or daily forecasting. Piggott [13] forecasts both two-hourly and weekly gas demands employing Box-Jenkins models which use weather input variables. For the two-hourly model he combines temperature and wind speed to calculate a chill factor. However when forecasting weekly gas demand for a few weeks ahead and no forecasts of temperature or wind speed are available, he excludes the wind speed variable and uses seasonal normal temperatures (SNT) for the weekly temperature data.

The above works deal with short term or peak load forecasting where regression models are employed for the modelling process. In this paper ARMA, ARMAX, SARIMA and Box–Jenkins transfer function models are used to carry out medium term forecasting. As in the above papers a weather variable, average temperature, is employed by forecasting models used in this paper; however this work considers the optimal configuration of the temperature input, depending on the data available. In the above cases of short term or peak load forecasting the weather variables used are considered to be dominant inputs. This paper studies two systems where average temperature is a dominant input, and where average temperature is a non-dominant input. The advantage of using a temperature input even when it is not a dominant input is considered. Only average temperature is considered here, some of the above works consider the use of wind variables and illumination variables which may be significant in medium term forecasting of electricity demand, perhaps these could
be considered in future work. Other works use Box–Jenkins, ARMA and ARMAX modelling techniques to forecast electricity demand, such as [7,9,10], each of these papers present in detail the steps taken to build these models and analyse the accuracy of the forecasting results.

2. Time series data

Weekly electricity consumption (KWh) and average temperature (°C) data are available from 4th April 1982 to 28th December 1991, a total of 508 points, and from the 5th April 1980 to 25th August 1990, a total of 543 points for the Irish data set and the New Zealand data set respectively. The average temperature values can be used to evaluate a heating degree day (HDD) figure which can then be accumulated to provide a weekly HDD figure. For example, weekly HDD values can be approximately evaluated using daily average temperature as follows,

If for each day, daily average temperature < base temp then

\[ \text{HDD}_{\text{daily contribution}} = \text{base temp} - \text{daily average temp} \quad (1) \]

else

\[ \text{HDD}_{\text{daily contribution}} = 0 \quad (2) \]

endif

\[ \text{HDD}_{\text{weekly}} = \sum_{\text{day } 1}^{\text{day } 7} \text{HDD}_{\text{daily contribution}} \quad (3) \]

The Irish and New Zealand data sets are presented in Figs. 1 and 2; note that the data has been scaled for confidentiality reasons. Figs. 3 and 4 show one year’s data from each case study, the effect of temperature on electricity demand can be seen more clearly in these graphs.

It can be seen that the consumption data have cyclical variations; for the Irish case these can be attributed to the seasonal temperature patterns throughout a given year. For the New Zealand case they are due to annual temperature profile variations but are also due to the cyclical consumption patterns of the seasonal industries (primarily the dairy industry). This is more obvious in Figs. 3 and 4. For the New Zealand data, the temperature decrease in the winter months has the effect of increasing the electricity consumption. However the demand continues to increase at a rate greater than the temperature decrease, this is due to the dominant influence of the agricultural industry. The peak of electricity consumption (in August) is not co-incident with the trough in temperature (in July), instead this peak corresponds to the peak in dairy production which occurs in Spring, i.e. August–October. Temperature is therefore a non-dominant input to the system. For the Irish data temperature is clearly the dominant input to the system, where the demand increases as the temperature decreases in the winter months and the demand decreases in the summer months.
3. Model building

For each case study the model building process involves using a portion of the data points for the identification of suitable models to describe the system (456 points for the Irish data set and 491 for the New Zealand data set), and the remaining 52 (corresponding to a year's duration) are used for model validation and comparison of results.
3.1. Box–Jenkins models

3.1.1. SARIMA univariate model

The univariate Box–Jenkins model is derived from the general SARIMA($p, d, q$)($P, D, Q$) (seasonal autoregressive integrated moving average) model.
written

$$\Phi_p(B)\Phi_p(B^L)\nabla Y_t = \delta + \Theta_q(B)\Theta_Q(B^L)a_t,$$

where

- $Y_t$ is the time series,
- $\nabla Y_t = (1 - B^d)(1 - B^d)$ is a differencing transformation required if the data is nonstationary, $d$ is the degree of non-seasonal differencing, $D$ is the degree of seasonal differencing, and $L$ is the number of seasons in a year,
- $a_t$ is the forecast error,
- $\Phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$ is the non-seasonal autoregressive operator of order $p$,
- $\Theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$ is the non-seasonal moving average operator of order $q$,
- $\Phi_p(B^L) = (1 - \phi_{1,L} B^L - \phi_{2,L} B^{2L} - \cdots - \phi_{P,L} B^{PL})$ is the seasonal autoregressive operator of order $P$,
- $\Theta_Q(B^L) = (1 - \theta_{1,L} B^L - \theta_{2,L} B^{2L} - \cdots - \theta_{Q,L} B^{QL})$ is the seasonal moving average operator of order $Q$,
- $\delta = \mu \Phi_p(B)\Theta_p(B^L)$ is the constant term, where $\mu$ is the true mean of the stationary time series being modelled.

The aim is to identify a suitable model structure for the time series in question based on the general form of the SARIMA model. To do this the orders $p$, $P$, $q$ and $Q$ of the $\Phi_p$, $\Phi_p$, $\Theta_q$ and $\Theta_Q$ operators must be chosen. A plot of the partial autocorrelation and autocorrelation function for the time series is examined in order to identify the lag where the functions “cut off” (i.e. is not significantly different from zero). The partial autocorrelation function and autocorrelation function are used to identify the orders of the autoregressive and moving average components respectively. The “cut off” at the non-seasonal level provide the values of $p$ and $q$ and at the seasonal level the values of $P$ and $Q$. The non seasonal level is at lags 1 through to $(L - 3)$ and the seasonal level is at lags approximately equal to $L$, $2L$, $3L$ and $4L$. Table 2 summarises this information.

Once the order of the model is selected the parameters are estimated; a popular method is the iterative least squares approach. The parameter estimates can be checked to determine if they are significant using the t-ratio test

$$t\text{-ratio} = \frac{\text{parameter estimate}}{\text{estimated standard error}}.$$  

<table>
<thead>
<tr>
<th>Autoregressive component</th>
<th>Moving average component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plot used</td>
<td>Partial autocorrelation function</td>
</tr>
<tr>
<td>“Cut off” at non-seasonal level after lag $p$</td>
<td>$p$</td>
</tr>
<tr>
<td>“Cut off at seasonal level after lag $p$</td>
<td>$p$</td>
</tr>
</tbody>
</table>
If the t-ratio lies outside the limits ±1.96, then the estimation of the parameter is assumed to be significantly different from zero and should be included in the model. There are a number of methods which may be used to test the adequacy of a model such as the Lung Box Test and the Durbin Watson Statistic. These methods are described in a textbook by Bowerman and O'Connell [3]. The identification, estimation and adequacy testing steps are iterated upon until an adequate model is found for the system.

3.1.2. Transfer function causal model

The general Box-Jenkins transfer function model is of the form

\[ z^{y(t)} = C \frac{\Omega(B)}{\Lambda(B)} z^{x(t-b)} + \eta(t), \]

(6)

where \( z^{y(t)} \) represents the stationary \( y(t) \) values, and \( z^{x(t)} \) represents the stationary \( x(t) \) values; \( b \) is a pure delay parameter, which represents the number of sampling periods before a change in the input series begins to have an effect on the output series.

\[ \Omega(B) = (1 - \omega_1 B - \cdots - \omega_s B^s) \]

where \( s \) is the number of past input values influencing current output values.

\[ \Lambda(B) = (1 - \delta_1 B - \cdots - \delta_r B^r) \]

and \( r \) is the number of past output values influencing current output values.

\[ \eta(t) \]

is a coloured (non-white) noise series of the form \( \mu + (\Theta(B)/\Phi(B))\psi(t) \).

A plot of the cross-correlation function (CCF) between the input series \( x(t) \) and the output series \( y(t) \) is used to evaluate the response lag time \( b \), and the orders \( r \) and \( s \) of the polynomials \( \Omega(B) \) and \( \Lambda(B) \). A "spike" (i.e. significantly different from zero) at lag \( k \) on the CCF plot indicates that \( x(t) \) is related to \( y(t - k) \). The lag where the first "spike" appears is set equal to \( b \). The CCF plot starts to decay a number of lags after lag \( b \), \( s \) is set equal to the number of lags that reside between lag \( b \) and the lag at which the decay begins. The values of \( r \) is determined by examining the manner in which the CCF plot decays. If the CCF decays in a damped exponential fashion, set \( r = 1 \), and if the CCF decays in a damped sine wave fashion set \( r = 2 \). Once the \( b, r \) and \( s \) values have been determined a preliminary transfer function model can be estimated. The residuals of the preliminary model are examined in order to identify a SARIMA model for the noise series \( \eta(t) \). The identification, estimation and diagnostic checking techniques described in Section 3.1.1 are used to determine the structure of the SARIMA noise model.

For more details of Box-Jenkins univariate and causal model building refer to books by Box and Jenkins [4], Pankratz [12] and Abraham and Ledolter [1].

3.2. ARMA and ARMAX models

3.2.1. ARMA univariate model

The univariate ARMA model is derived from the general parametric model structure for describing a system based on observed input–output data written

\[ A(q)y(t) = \frac{B(q)}{F(q)} u(t - nk) + \frac{C(q)}{D(q)} e(t), \]

(7)
where \(u(t), t = 1, 2, \ldots, N\), is the input signal, \(y(t), t = 1, 2, \ldots, N\), is the output signal, \(e(t)\) is white noise.

The \(A, B, C, D\) and \(F\) are polynomials in the delay operator \(q^{-1}\):

\[
A(q) = 1 + a_1 q^{-1} + \cdots + a_{na} q^{-na} \\
B(q) = b_1 + b_2 q^{-1} + \cdots + a_{nb} q^{-nb} \\
C(q) = 1 + c_1 q^{-1} + \cdots + c_{nc} q^{-nc} \\
D(q) = 1 + d_1 q^{-1} + \cdots + a_{nd} q^{-nd} \\
F(q) = 1 + a_1 q^{-1} + \cdots + a_{nf} q^{-nf}
\]

\(na, nb, nc, nd\) and \(nf\) are the orders of the polynomials, \(nk\) is the number of delays from input to output.

The ARMA \((na nc)\) model is obtained when \(nb = nd = nf = 0\),

\[
A(q)y(t) = C(q)e(t).
\]

The orders of the \(A(q)\) and \(C(q)\) polynomials \((na\) and \(nc)\) are determined by plotting and comparing the loss function for different values of \(na\) and \(nc\). The loss function provides a measure of the mean square difference between the model output and the actual consumption for a particular model structure for zero initial conditions. The objective is to pick polynomial orders which are just sufficient to describe the system. A systematic way to do this is to employ Akaike's Information Theoretic Criterion (AIC) which is order weighted, i.e. it penalises higher orders. The AIC is given by

\[
AIC \approx \log\left(\frac{1 + 2n}{N} \ast V\right).
\]

\(n\) is the total number of estimated parameters, \(N\) is the length of the data record, \(V\) is the loss function for the structure in question, where in a collection of models with different structures the one with the smallest AIC should be chosen; refer to Ljung [8] for examples. Once the model order is chosen the parameters are estimated through the use of an efficient estimation method such as the least squares approach.

3.2.2. ARMAX causal model

Using (7) an \(\text{ARMAX}(na \ nb \ nc)\) model is defined for \(nd = nf = 0\),

\[
A(q)y(t) = B(q)u(t) + C(q)e(t).
\]

The orders of the \(A(q), B(q)\) and \(C(q)\) polynomials \((na, nb\) and \(nc)\) are determined and estimated using the methods described above in Section 3.1.2.

4. Selection and evaluation of system temperature inputs

As discussed in Section 2 average temperature (°C) or HDD data is available for use as inputs to the causal electricity forecasting models. For each case study it is
required to select the temperature input which has the strongest correlation with electricity consumption. It is first necessary to calculate the most accurate HDD values.

4.1. Calculation of HDD

The accurate calculation of the weekly HDD values requires the determination of the following.

(i) The correct HDD base temperature. It is desirable to select for each data set the base temperature which yields the weekly HDD data series which is the most highly correlated with the electricity consumption data series. The base temperature may vary between each case study due to the basic differences between each system but also due to the different comfort levels of the people in each country. The correlation between weekly electricity consumption and weekly HDD using base temperatures in the interval 14°C–12°C are compared.

(ii) The time scale upon which to base the HDD calculations. Having selected the HDD base temperature, it is necessary to determine the time scale on which to base the average temperature values to use in the calculation of the weekly HDD. If hourly average temperature is available there are three possible methods of calculating HDD.

Method 1: Use hourly average temperature data to calculate weekly HDD as follows:

If for each hour, hourly average temperature < base temp then

\[
\text{HDD}_{\text{hourly contribution}} = (\text{base temp} - \text{hourly average temp})
\]  \hspace{1cm} (12)

else

\[
\text{HDD}_{\text{hourly contribution}} = 0
\]  \hspace{1cm} (13)

endif

\[
\text{HDD}_{\text{daily contribution}} = \frac{1}{24} \int_{00:00}^{23:00} \text{HDD}_{\text{hourly contribution}}
\]  \hspace{1cm} (14)

\[
\text{HDD}_{\text{weekly contribution}} = \int_{\text{day 1}}^{\text{day 7}} \text{HDD}_{\text{daily contribution}}
\]  \hspace{1cm} (15)

Method 2: For each day, calculate a daily average temperature value using the hourly data. Use the daily average temperature values to calculate weekly HDD as follows:

If for each day, daily average temperature < base temp then

\[
\text{HDD}_{\text{daily contribution}} = \text{base temp} - \text{daily average temp}
\]  \hspace{1cm} (16)
Method 3: For each week, calculate a weekly average temperature value using the hourly data. Use the weekly average temperature values to calculate weekly HDD as follows:

If for each week, weekly average temperature < base temp then

\[ \text{HDD}_{\text{weekly}} = \left[ (\text{base temp} - \text{weekly average temp}) \times 7 \right] \]

else

\[ \text{HDD}_{\text{weekly}} = 0 \]

endif

Each of the above methods produces different weekly HDD figures. The final weekly HDD data differs significantly during the summer months, i.e. May–September for the Irish data set and November–April for the New Zealand data set; the HDD figures calculated using hourly average temperature values are the greatest in magnitude. For the remaining months of the year the differences between the HDD figures are minor. The reason for this is because for a typical winter day the hourly average temperature would rarely rise above the HDD base temperature, therefore taking an average of these hourly values (i.e. method 2 and method 3) and the subsequent calculation of weekly HDD does not result in as great a loss of the HDD contribution as with the summer data where it is expected that the hourly average temperature would rise above the HDD base temperature at certain times of the day. It is necessary to test whether any of the above methods produce weekly HDD figures which are more highly correlated with the electricity consumption data than the others.

4.2. Mathematical methods for correlation analysis

The selection of the best temperature input to use for the causal forecasting models, and the determination of the most appropriate HDD calculation method each require a method of estimating the degree of correlation between the input and output data series. Two such methods are now described.

4.2.1. Correlation coefficient

Given two series \( X_n \) and \( Y_n \), the covariance between the series is calculated as

\[ \text{Cov}_{xy} = \frac{1}{N} \sum_{n=0}^{N} (x_n - \bar{x})(y_n - \bar{y}). \]
The correlation coefficient \( r_{xy} \) is the normalised version of the covariance and is defined as

\[
r_{xy} = \frac{1}{N} \frac{\sum_{n=0}^{N} (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_{n=0}^{N} (x_n - \bar{x})^2 \sum_{n=0}^{N} (y_n - \bar{y})^2}}.
\]  

(22)

The correlation coefficient estimates the degree of closeness of linear relationship between the two variables \( X \) and \( Y \). The value of \( r \) lies on the interval \(-1 \leq r \leq 1\). It is zero when the two variables are independent. When \(|r_{xy}| \approx 1\), \( X \) and \( Y \) are said to be highly correlated. The correlation coefficient can be used to some extent to measure the association between two variables; however due to the fact that it treats the variables symmetrically it does not provide any information on the cause and effect relationship which may exist between these variables. It is possible to calculate confidence limits about \( r \). When testing for \( r \neq 0 \), i.e. for correlation, it is necessary to apply a transformation to convert \( r \) to a quantity \( z \) which is distributed normally with a standard error of approximately

\[
\sigma_z = \frac{1}{\sqrt{n - 3}},
\]

(23)

where \( n \) is the number of observations. The relation of \( z \) to \( r \) is given by

\[
z = \frac{1}{2}[\ln(1 + r) - \ln(1 - r)].
\]

(24)

Refer to the books by Snedecor and Cochran [16] and Morrison [11] for more details concerning this transformation and the construction of confidence limits. If confidence intervals are constructed around each of the \( r \) values comparisons can be made based on the property that the closer \(|r_{xy}|\) is to 1 the higher the correlation between the variables \( X \) and \( Y \).

4.2.2. Cross covariance function

Given two stationary series \( X \) and \( Y \); with \( E[X] = 0 \) and \( E[Y] = 0 \) the cross covariance function between the series is calculated as

\[
R_{xy}(\tau) = \frac{1}{n} \sum_{t=\tau}^{n-\tau} x(t)y(t + \tau),
\]

(25)

where \( n \) is the number of data points in each series. The function provides a measure of the association between two variables at a response lag \( \tau \). A normalised form or the equation is given as

\[
r_{xy}(\tau) = \frac{R_{xy}(\tau)}{\sqrt{R_{xx}(0)R_{yy}(0)}}.
\]

(26)

\( r_{xy}(\tau) \) can be calculated for different values of \( \tau = 0, 1, 2, \ldots \) and a plot of \( r_{xy}(\tau) \) as a function of \( \tau \) can be obtained. Correlation between the two series exist at the values of \( \tau \) where \( r_{xy}(\tau) \) is significant. It is possible to calculate confidence limits for the cross covariance function as follows.
Assume that $X$ and $Y$ are not related and independent then,

$$\sqrt{n}R_X(\tau) \in N(0, \sigma),$$  \hspace{1cm} (27)

$$P = \sum_{k=-\infty}^{\infty} R_X(k)R_Y(k),$$  \hspace{1cm} (28)

$$R_X(k) = E[X(t)X(t-k)] \quad \text{and} \quad R_Y(k) = E[Y(t)Y(t-k)].$$  \hspace{1cm} (29)

For an $\alpha\%$ confidence limit, let $N_\alpha$ denote the $(1-\alpha)$ level of the $N(0, 1)$ distribution. Then to check if $X$ and $Y$ are related check if the following is true,

$$|R_{XY}(\tau)| \geq \sqrt{\frac{P}{n}}N_\alpha.$$  \hspace{1cm} (30)

Since $P$ does not depend on $\tau$ the confidence limits will be horizontal lines and can be plotted on the same graph as $r_{xy}(\tau)$. If the cross covariance function lies outside of this horizontal line at lag $\tau$ then there is an $\alpha\%$ chance that the variables are correlated at lag $\tau$. Refer to books by Ljung [8] and Soderstrom and Stoica [17] for more details.

The correlation coefficient is a measure of the closeness of linear relationship between two variables but it does not give any insight into the cause and effect relationship as does the cross covariance function. The coefficient is straightforward to calculate and can be used to get an initial estimate of the linear association between input and output variables. When using a temperature input to an electricity consumption model it is expected for most systems that the lag between consumption and temperature will be zero. However for long term electricity demand forecasting economic variables are often significant inputs and in this case the cross covariance function can be used to determine the time lag between input and output variables. For the systems under-study in this paper the cross covariance function is used with confidence limits to measure if one set of HDD values are more highly correlated with electricity consumption than another at lag zero. In general it is suggested that both tests be applied to determine the form of the relationship between two variables.

4.3. Application of correlation analysis

The above methods of measuring the degree of association between two data series are applied to the following

(i) selection of base temperature to use when calculating weekly HDD figures,

(ii) selection of time scale to use when calculating weekly HDD,

(iii) selection of temperature input (average temperature or HDD) which has the highest degree of correlation with electricity consumption.

(i) **Selection of HDD base temperature.** The correlation coefficient and cross covariance function of electricity consumption with each of the weekly HDD data series calculated using the basic temperatures in the interval 14°C–21°C are computed. The results are compared where the objective is to select the base temperature which yields the HDD series which has the highest degree of correlation with electricity
consumption. For both the New Zealand and Irish case study a base temperature of 18°C is selected. The HDD values are correlated with electricity demand at lag zero. The correlation coefficient results are given in Table 3. The correlation coefficient values are higher for the Irish data than for the New Zealand data as expected given that HDD is a dominant input.

(ii) Selection of time scale to use when calculating weekly HDD. For the New Zealand case study only daily average temperature values are available therefore methods 2 and 3 are used to calculate the weekly HDD figures. For the Irish case study hourly average temperature data is available and the weekly HDD figures can be calculated using methods 1, 2 and 3. The figures produced by each method differ significantly during the summer months only, therefore the correlation coefficient between electricity consumption and each of the weekly HDD data series calculated using the different methods are computed for the summer data and compared. Using the cross correlation coefficient as a criteria no one method produces weekly HDD figures which are more highly correlated with electricity consumption. For each case study causal models were built using the different weekly HDD figures as inputs. The accuracy of the predictions resulting from the simulations of these models are compared using the Mean Absolute Fit (MAF); where the MAF is calculated as

$$MAF = \frac{\sum_i |Y_P^i - Y_A^i|}{n}$$

where $Y_P$ is the $i$th prediction, $Y_A$ is the corresponding $i$th actual value and $n$ is the number of predictions made in the forecast simulation. Using the MAF as a basis, the best result for the New Zealand data set was obtained using the HDD input figure calculated using daily average temperature. The Irish data set calculating the HDD input figures using hourly average temperature values yielded the most accurate results. Table 4 gives the MAF results.

(iii) Selection of temperature input (average temperature or HDD) which has the highest degree of correlation with electricity consumption. For both the Irish and New Zealand data sets, the correlation coefficient and cross covariance function of weekly

<table>
<thead>
<tr>
<th>Table 3</th>
<th>HDD base</th>
<th>14°C</th>
<th>15°C</th>
<th>16°C</th>
<th>17°C</th>
<th>18°C</th>
<th>19°C</th>
<th>20°C</th>
<th>21°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irish</td>
<td>Correlation coefficient</td>
<td>0.85</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.87</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>New Zealand</td>
<td>Correlation coefficient</td>
<td>0.49</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.51</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4</th>
<th>MAF – hourly average temp</th>
<th>MAF – daily average temp</th>
<th>MAF – weekly average temp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irish</td>
<td>0.89 e + 04</td>
<td>0.93 e + 04</td>
<td>0.94 e + 04</td>
</tr>
<tr>
<td>New Zealand</td>
<td>N/A</td>
<td>2.70 e + 05</td>
<td>2.95 e + 05</td>
</tr>
</tbody>
</table>
electricity consumption with weekly average temperature, and weekly electricity consumption with weekly HDD are computed and used to compare the degree of the correlation between the input and output variables. For each case study the HDD data is the most highly correlated with the consumption data and is thus chosen as the input to the causal forecasting models. Table 5 gives the correlation coefficient results.

5. Simulation results

The Box–Jenkins SARIMA and transfer function models are built and simulated using BMDP software, (BMDP 1988), on a DECVAX 6230. The ARMA and ARMAX models are identified and simulated using the MATLAB package with the System Identification Toolbox (Matlab for Microsoft Windows 1991) run on a 386 personal computer. The simulations predict weekly electricity consumption 52 weeks in advance. For each data set the last 52 data points were reserved for model validation, and also the comparison of the accuracy of the predictions obtained from the univariate and causal models. The MAF is computed for each set of predictions.

Table 5

<table>
<thead>
<tr>
<th></th>
<th>HDD\textsubscript{18}</th>
<th>Average temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irish</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>New Zealand</td>
<td>0.51</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table 6
Box–Jenkins SARIMA and transfer function models

<table>
<thead>
<tr>
<th>Univariate model</th>
<th>Causal model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Structure</td>
</tr>
<tr>
<td>SAR p = (1, 2, 3)</td>
<td>P = (52, 104)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7
ARMA and ARMAX models

<table>
<thead>
<tr>
<th>Univariate model</th>
<th>Causal model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Structure</td>
</tr>
<tr>
<td>ARMA na = 53, nc = 5</td>
<td>3.39 e + 05</td>
</tr>
</tbody>
</table>
|                  |                      |           | ARMAX HDDpred  | na = 53, nb = 2, nc = 1 | 3.213 + 05
As part of the model validation process the causal models are simulated using actual and predicted temperature inputs. The MAF values for the New Zealand case study simulations are presented in Tables 6 and 7 and for the Irish case study in Tables 8 and 9. Note $\eta_t$ represents the noise model for the Box-Jenkins transfer function model. Examples of the forecasting results obtained for both case studies are given in Figs. 5–10.

Table 8
Box-Jenkins SARIMA and transfer function models

<table>
<thead>
<tr>
<th>Univariate model</th>
<th>Causal model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td><strong>Structure</strong></td>
</tr>
<tr>
<td>SAR</td>
<td>$p = (1, 2, 3, 4, 5, 6)$</td>
</tr>
<tr>
<td></td>
<td>$P = (52, 104)$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9
ARMA and ARMAX models

<table>
<thead>
<tr>
<th>Univariate model</th>
<th>Causal model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td><strong>Structure</strong></td>
</tr>
<tr>
<td>ARMA</td>
<td>$na = 53, nc = 1$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>ARMAX</td>
<td>HDDpred</td>
</tr>
</tbody>
</table>

Fig. 5. Actual vs. predicted electricity demand. New Zealand data – using no input to the forecasting model.
6. Conclusions

This paper shows that an improvement in the accuracy of electricity consumption forecasting results may be achieved through the use of an appropriately selected input; this was shown to be true for both dominant and non-dominant inputs. The degree of improvement obtained is greater for the dominant input case. This increased accuracy is attained for both cases even when future values of the exogenous input are unknown and it is necessary to predict it through the use of a univariate forecasting model. The relative order of magnitude of the autoregressive and exogenous parameters of the causal forecasting model indicate that the exogenous input provides a supplement adjustment to the forecast obtained from the pure univariate autoregressive model (the autoregressive and temperature input contributions to the electricity demand forecast are of the order of $10^5$ and $10^3$ respectively for the Irish

![Fig. 6. Actual vs. predicted electricity demand. Irish data - using no input to the forecasting model.](image)

![Fig. 7. Actual vs. predicted electricity demand. New Zealand data - using HDD actual as input to the forecasting model.](image)
The paper considers methods of determining the optimal configuration of the temperature input for the electricity demand systems. For both case studies, it was found that HDD was a more effective input than average temperature and a base temperature of 18°C was selected in each case. This was verified both using correlation analysis and forecast accuracy. The evaluation of HDD was also studied with respect to the availability of temperature data on various time scales. It was not possible to use correlation analysis to ascertain the best time scale to use when calculating HDD, however using the Mean Absolute Fit it was found for each case study that the use of the finest average temperature time scale available yielded the HDD values which had the highest correlation with electricity consumption.
Fig. 10. Actual vs. predicted weekly HDD. Irish Data.

References