Pak J. Statist.
2003 Vol. 19(3) pp 353 – 373

MODEL USED TO DETERMINE THE DAILY AVERAGE DEMAND OF ELECTRIC ENERGY IN ARGENTINA – A STATE SPACE APPROACH

María Teresa Blaconá
National University of Rosario
Boulevard Oroño 1261, 2000 Rosario, Argentina
mblacona@agatha.unr.edu.ar

and

Juan Carlos Abril
National University of Tucumán and CONICET,
4000 Tucumán, Argentina
jabril@herrera.unt.edu.ar

ABSTRACT

This work shows the usefulness of state-space models to adjust and forecast daily time series, and the technique of periodic cubic spline regression to model annual seasonality. A structural model is used to analyzed the series of daily average demand of electricity in Argentina. This model considers the trend, the weekly and annual seasonal component, the effect of public holidays, two cycles, and the temperature as explanatory variable. The method gave satisfactory results, both at the adjustment level as well as in the forecasting and interpretability of its components. Alternative methods are recommended when the future temperature values are unknown.

KEY WORDS

Daily time series, electricity demand, Kalman filter and smoothing, periodic cubic spline, state space approach, structural model.

1. INTRODUCTION

The Wholesaler Electrical Market (MEN in Spanish) of our country happens to be regulated by law. According to it the MEN is the real place where you can find both the supply and the demand of electric energy in our country. All the energy generators must be MEN's agents just like the firms that act in the market, with the only exception of small consumption. These agents provide all the information that is needed to create the database that belongs to the Wholesaler Electrical Market Administrator Company (Cammesa in Spanish). This agency builds up such a database and makes it available to such agents. More than 90% of the electric energy demand is provided by the MEN, being Cammesa the one assigned to determine which power station will operate, besides supervising a few items.

One of the most important goals of any electrical system is supplying the market with the least interruptions and keeping the quality of the energy offered. Therefore, it is necessary to have an efficient system.

For all these reasons, this paper tries to find a parsimonious model in order to:

a) Describe the behaviour of the daily average demand series of electrical energy for Argentina.
b) Carry out short run predictions.

To do this we use tools such as the structural models and the state space approach.

The main drawback we find in modelling daily time series is fitting the seasonal component, due to the fact that we have two kinds of seasonality: weekly and annual. In the first case, it is enough to use a basic structural model where the seasonal component is modelled in trigonometric terms or we can use "dummy" variables. In the second case, the problem we have is much more complex, because traditional models would require a larger number of parameters, not fulfilling the parsimony principle.

To solve this problem we propose to use structural time series models (see Harvey, Koopman and Riani, 1996) which can be interpreted as regressions over functions of time with varying parameters. This allows us to treat seasonal patterns that vary in a complex way. When we adjust a convenient model, the seasonal component can be estimated by some smoothing algorithm. In other words, we can generalize a deterministic component in order to get a stochastic one.

We use a highly recommended technique called "spline" (see Poirier, 1973) to treat the annual seasonality, because it is quite simple implement and because it permits a non-linear effect to change into a multiple lineal regression allowing the estimation. Besides, it is possible to work with periodical "splines" and make some restrictions that ensure all seasonal components to add up zero. In this way, there is no confusion between the seasonal and the trend components. In some cases, it is useful to add to the model one or more explanatory variables related to the phenomenon of our interest. In this case we include the daily average temperature series, to help explaining the behaviour of the daily average demand of energy.

Once the model is set up, it is relative easy the statistical handling under the state space approach. The characteristic of the space state form for modelling any system over time is that includes two very different stochastic processes. In one the distribution of the data at each point of time is conditional to a set of parameters indexed by time. A second process describes the evolution of the parameters over time.

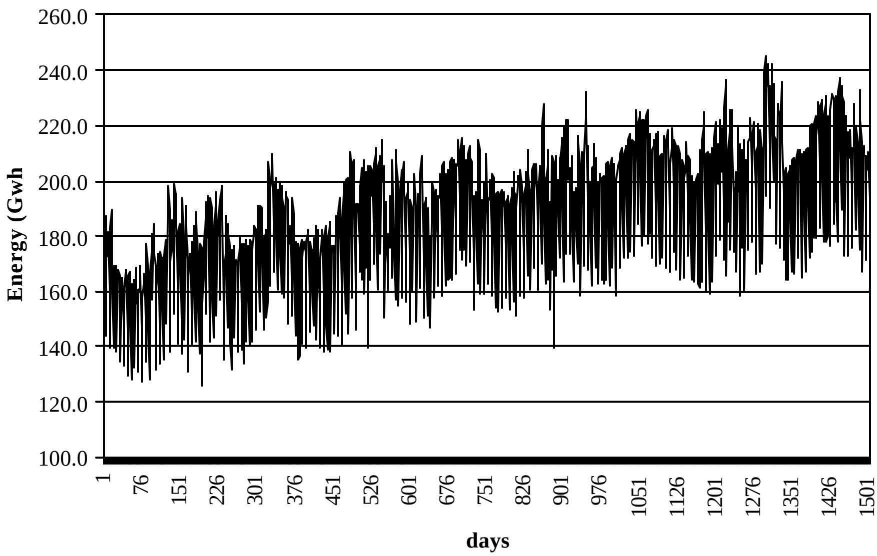
To sum up, forecasting daily time series is a hard task that has to be developed in a permanent way. The main characteristic of many daily time series is that the seasonal pattern changes persistently over time. State space modelling with its recursive estimation techniques and its statistic methodology is an attractive way to deal with the mentioned problem.

The proposed models are very similar to those used before by Harvey and Koopman (1993) in order to forecast the Puget Sound Power and Light (USA) hourly demand series and by Gordon (1996) in order to model the daily energy demand series by three Brazilian companies LIGHT, CEMIG and COPEL. Very good results of fitting and prediction were obtained in both works, which encouraged us to use a similar method in the case of Argentina. In Section II we make a brief description of the series. In Section III we introduce the state space models for the daily series. In Section IV we present the spline technique applied to daily time series. In Section V we introduce the complete model with its estimations, goodness–of–fit tests and forecasts. Finally, we conclude our discussion in Section VI.

2. CHARACTERISTICS OF THE SERIES

The information we have is the argentinian energy daily average demand measured in Gigawath (Gwh) for the period between August 1st 1995 and September 12th 1999 that was provided by Cammesa. We also have the daily average temperature series measured in degrees centigrade (°C) for the same period of time, provided by the same source.

Figure 1: Daily Average Demand of Energy in Argentina
(01-08-95 - 12-09-99)



The energy demand series presents a smooth growth and a clear seasonal behaviour (Figure 1). This last kind of behaviour is characterised by two very neat periodical parts: one annual (Figure 2) and the other weekly (Figure 3).

The annual seasonal component is due to a combination of factors, among which are: temperatures, wind, sunlight hours; economic factors just like productive cycle, price (this variable is not very important in the case of a daily measure) and demographic factors (this is also not very important when we treat daily series). These graphics show a very distinctive behaviour between the demand on working days (Monday to Friday) and on Saturdays (where the demand diminishes) and Sundays (which is the lowest demanding day). It is also remarkable that in holidays the demand diminishes and this reduction can vary according to the day we have a holiday (Figure 3).

Figure 2: Daily Average Demand of Energy in Argentina (1-8-98 - 31-7-99)

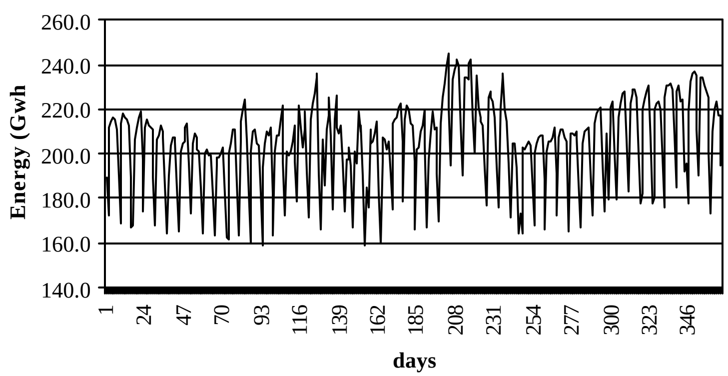
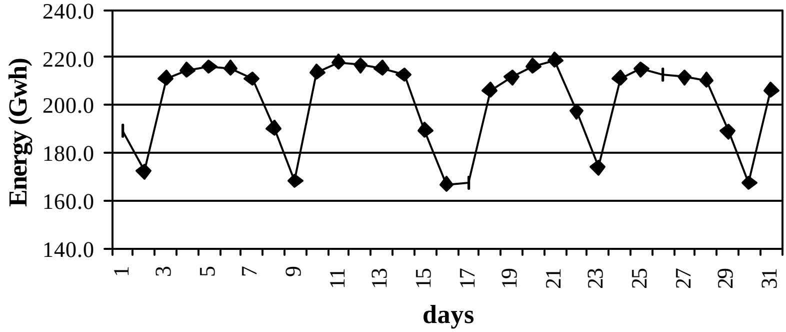


Figure 3: Daily Average Demand of Energy in Argentina. August 1998



All the stated in the previous paragraphs show us that it is a highly difficult task to find a parsimonious model (with few parameters) that capture all the behaviours present in the demand series.

3. STRUCTURAL MODELS

Structural models can be interpreted as regressions over functions of time in which the parameters are allow to vary in time. This makes them a natural vehicle to treat complex stationary behaviour.

3.1 Basic state space model

The basic state space model (BSSM), also known as Gaussian lineal space state model can be expressed as,

$$y_t = Z_t \alpha_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, H_t),$$
$$\alpha_t = T_t \alpha_{t-1} + R_t \eta_t, \quad \eta_t \sim NID(0, Q_t), \tag{1}$$

where y_t is a $p \times 1$ vector of observations, α_t is an $m \times 1$ vector called "state vector", the matrixes Z_t (of order $p \times m$), T_t (of order $m \times m$) and R_t (of order $m \times g$) are assumed known and the error terms ε_t (of order $p \times 1$) and η_t (of order $g \times 1$) are assumed serially independent and independent among them at all time moments. We introduce the matrix R_t as a selection matrix formed only by zeros and ones whether the corresponding element of η_t is random or deterministic.

The first equation in (1) is usually called the measurement equation and the second one, the transition equation. In the following section we will specify a structural model for daily time series.

3. 2 Basic structural model

Structural time series models are set up explicitly in terms of components that have a direct interpretation (see Abril, 1999 and Harvey, 1989). Suppose we have the following time series $y_1, \ldots, y_T$. The basic structural model (BSM) is formulated in terms of trend, seasonal and irregular components. The model can be expressed as,

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_t^2), \quad t = 1, \ldots, T, \tag{2}$$

where μ_t, γ_t and ε_t represent the trend, seasonal and irregular components respectively, and are in principle stochastic. In some cases the components combine multiplicatively, but taking logarithm and working with the logarithm of the series it is possible to obtain the additive form given in (2). The main idea in these models is that the trend and the seasonal components conform the principal aspects of the time series. That is why the construction of the model is oriented to the estimation and analysis of these components.

The specification of μ_t, γ_t and ε_t is based on the knowledge we have about the process we are analysing and on basic statistics techniques. The seasonality is usually represented by stochastic trigonometric functions in the seasonal frequencies $s/2$ (where s is the seasonal period), or by "dummy" variables. The key point is that even though the seasonal component is not stationary it has the property that the expected value of the sum over the s previous periods is zero. This ensures that the seasonal effect will not be confused with the trend and it also means that the forecasts made of the seasonal component must add up zero over any annual period. On the other hand, in classical models μ_t is generally defined as a polynomial in time, γ_t is specify using "dummy" variables or trigonometric functions and ε_t as an ARMA process (see Box and Jenkins, 1971). That is, the trend and the seasonal effects are supposed to be deterministic. In many situations these assumptions are not true, therefore the structural models are very useful.

3.3 Model specification for a daily energy series

As we mentioned earlier, the primary component in fitting daily time series is the seasonal, due to the fact that we have two kinds of seasonality:

I. Weekly II Annual

In the following section we will deal with point I, leaving point II for later treatment when we will use the "spline" technique in order to incorporate this component to the structural model. It is also important to have in mind the influence of holidays in the series. Besides, we can include explanatory variables to the model in order to understand the problem of our interest.

3.3.1. Weekly seasonality

In daily time series it is generally presented a weekly periodical behaviour $(s = 7)$ which can be treated in two ways: a) with "dummy" variables, or b) using trigonometric functions. In this work we focus on point a).

Suppose we have the seasonal parameters γ_j, $j = 1, \ldots, 7$. If we consider the seasonality to be deterministic, we have to say that the sum of all effects must be zero, that is,

$$\sum_{j=1}^{7} \gamma_j = 0, \tag{3}$$

or $\sum_{j=1}^{7} \gamma_{t-j} = 0$. In this case

$$\gamma_t = -\sum_{j=1}^{6} \gamma_{t-j} \quad t = 7, 8, \ldots. \tag{4}$$

On the other hand, there are many situations where the seasonality varies over time, in which case we say that it is stochastic. A simple way to deal with this is to add up an error term to the previous expression, that is,

$$\gamma_t = -\sum_{j=1}^{6} \gamma_{t-j} + w_t, \quad t = 1, \ldots, T, \quad w_t \sim NID(0, \sigma_w^2). \tag{5}$$

In this case, instead of considering the sum to be zero, we have to suppose that the expected value is zero, which make the "dummy" variables flexible to change over time.

3.3.2. Holidays

In daily time series there is usually a holiday effect that act upon the series. We can set a "dummy" variable to treat this situation. Holidays can be considered as an annual periodical component, whose effects must add up to zero every year. That is, "dummy" variables can be represented by θ_{ij}, $i = 1, \ldots, q$, where *i* represent the day or the set of days of the week where we can find the holiday, and $j = 1, \ldots, m$, where *j* represents the year.

η_{ij} is defined as the number of times that θ_{ij} can be found in year *j*. In order that the holiday effect has no influence on the level, it should satisfy

$$\sum_{i=1}^{q} \eta_{ij}\ \theta_{ij} = 0, \tag{6}$$

in consequence,

$$\theta_{qj} = -\frac{1}{\eta_{qj}} \sum_{i=1}^{q-1} \eta_{ij}\ \theta_{ij}, \tag{7}$$

Due to the kind of holidays we have in Argentina and its influence on energy consumption, we can define the following variables that measure the holiday effect:

θ_1 if the holiday is on Monday.

θ_2 if the holiday is from Tuesday to Friday.

θ_3 if the holiday is on Saturday.

θ_4 if the holiday is on Holy Thursday.

θ_5 if there is no holiday.

Then

$$\sum_{i=1}^{5} \eta_{ij}\ \theta_{ij} = 0,$$

in consequence,

$$\theta_{5j} = -\frac{1}{\eta_{5j}} \sum_{i=1}^{4} \eta_{ij}\ \theta_{ij}\ .$$

3.3.3 Explanatory variables

Explanatory variables and intervention effects can easily be incorporated into structural models. For instance, if we suppose that there are k regressors $x_{1t}, \ldots, x_{kt}$, they can be incorporated directly into the equation of the model, adding the term $\sum_{j=1}^{k} \delta_j\ x_{jt}$ to the equation (2) of the structural model. We have to act on a similar fashion if we add intervention variables because of the presence of "outliers" o structural changes. In the case of the energy demand there is a quadratic relation with the temperature, because the energy consumption rises on very cold days as well as on hot days. Regarding the variables that measure the productive cycle, we cannot study any specific relation with the energy demand because there is no daily record of them. The price variables do not have much effect on the daily energy demand since in these cases the changes are imperceptible.

3.4 Estimation of unobservable components

In the previous section a state space model was proposed, where unobservable components intervene in the state system. The next step consists in the estimation of those components. The whole process is as follows,

I. Filtering: it is an operation where the system is updated every time a new observation y_t is available. This filtering process is made using the Kalman filter technique.
II. Initiation: We must specify how we start the filtering process, that is, we specify the initial values of the components.
III. Smoothing: This considers the estimation of the components using the complete sample information.

The smoothing and the filtering processes must be done at the same time. Two set of estimates are obtained by these two processes, the very best by Fisher's point of view is the smoothed set because it uses all the information in the sample.

3.5 Estimation of the hyper-parameter by maximum likelihood

The parameters in state space models are usually called hyper-parameters probably to distinguish them of from the elements of the state vector that can be considered as stochastic parameters. Suppose we have the hyper-parameter vector φ, that is going to be estimated using maximum likelihood (ML). In time series the observations are not independent, therefore for the construction of the likelihood we use properties of the conditional distribution like the following,

$$L(Y,\varphi)=\prod_{t=1}^{T} p\left(y_t / Y_{t-1},\varphi\right), \tag{8}$$

where $p\left(y_t / Y_{t-1}\right) \sim N\left(Z_t,\alpha_t,F_t\right)$ and Y_{t-1} is the information up to moment $t-1$. So if we take logarithm we have,

$$\log L=-\frac{T\,p}{2}\log(2\pi)-\frac{1}{2}\sum_{t=1}^{T}\log\left|F_t\right|-\sum_{t=1}^{T} v_t' F_t^{-1} v_t\,, \tag{9}$$

where $v_t = y_t - Z_t\, a_t$ is the prediction error distributed as $NID\left(0,\, F_t\right)$. This last equation is named prediction error decomposition of the log likelihood and it must be maximized with respect to the elements in the vector φ of unknown hyper-parameters.

When a_0 and P_0 are not known, it is supposed that τ is the lowest value of t for which $p\left(\alpha_t / Y_t\right)$ exists. Then, we take the conditional likelihood with Y_τ fixed and obtain,

$$\log L=-\frac{(T-\tau)\,p}{2}\log(2\pi)-\frac{1}{2}\sum_{t=\tau+1}^{T}\log\left|\mathrm{F_t}\right|-\sum_{t=\tau+1}^{T} v_t' F_t^{-1}\, v_t\,, \tag{10}$$

Expressions (9) and (10) are maximized by a maximization algorithm. The optimal method used by STAMP 5.0 (software employed in this work) is based on a method called BFGS Quasi-Newton (see Koopman, Harvey, Doornik and Shephard, 1995). The probes we make to see if a certain model is correct are based on the estimated standarized innovations $\hat{v}_t \big/ \hat{F}_t^{-1/2}, \quad t=d+1,\ldots,T$ (see Koopman, Harvey, Doomik and Shephard, 1995) calculated from the smoothed residual $\hat{v}_t$. If the model is well specified, these residuals have the advantage of being approximately non-correlated.

The basic measure of the goodness of fit is called PEV, that is *the prediction error variance*, or $\tilde{\sigma}^2$. Another measure of the goodness of fit is the *average deviation of the residuals* which is defined as,

$$md=\frac{\tilde{\sigma}}{T-d}\sum_{t=d+1}^{T}\left|v_t\right|,$$

and, if the model is well specified the ratio,

$$c = \frac{2\tilde{\sigma}^2}{\pi(md)^2}$$

is close to one.

There are many techniques to probe the goodness of a certain model. We will display those that help us see if the model is satisfactory.

- Heteroscedasticity: It is based on a non-parametric bilateral test of heteroscedasticity, F_{h,h_1}, where we build the statistic,

$$H(h) = \sum_{t=T-h+1}^{T} v_t^2 \Big/ \sum_{t=d+1}^{d+1+h} v_t^2 ,$$

being h the closest integer to $(T - d)/3$.

- Residual correlation: There are basically two kinds of statistics to probe if the residuals are correlated:

I. Durbin-Watson statistic: This help us probe if the residuals present a first order autocorrelation, and it is given by

$$DW = \sum_{t=2+d}^{T} (v_t - v_{t-1})^2 \cong 2(1 - r(1));$$

if the model is well specified DW is distributed as $N(2, 4/T)$.

II. Box- Ljung Q statistic: this allows us to probe if the first P autocorrelations are equal to zero and it is given by

$$Q(P,k) = T(T+2)\sum_{j=1}^{P} r_j^2 /(T-j),$$

under the hypothesis that the residuals are not correlated it is distributed as c_k^2, where $k = P - n + 1$, and n is the number of hyper -parameters.

A useful measure to check how good the fitting is happens to be the *determination coefficient*. In the case when we analyse seasonal series with a trend, like daily time series, we use an adjusted coefficient given by

$$R_s = \frac{1-(T-d)\tilde{s}^2}{SSDSM}$$

where $SSDSM$ is the sum of square obtained deducting the seasonal mean to the first difference of y_t.

To compare models with different number of parameters the PEV is not convenient. A more proper coefficient called *Akaike Information Criterion* (AIC), which penalize the PEV by the number of estimated parameters on the model is defined as,

$$AIC = \log PEV + 2m / T ,$$

where m is the number of hyper-parameters plus the number of unobservable parameters. In consequence, the lowest the value the AIC the better is the fitting.

Besides all the goodness of fit measures we mentioned, it is important to remind that:

I. The estimation of the hyper-parameter allows us to know if the components move in a stochastic way. Nevertheless, a zero value of hyper-parameter indicates a deterministic behaviour.

II. The graphic of the smoothed components helps us to see if the decomposition made by and adjusted model was useful. In terms of prediction the estimated trend is the part of the series that indicates a future movement in the long run (see Harvey, 1989), for that it is important that it does not have a periodical behaviour.

3.6 Prediction

In the Gaussian model defined in (1), the Kalman filter leads to a_T, the minimum quadratic error estimator of α_T, based on all the observations. Besides, it is known that,

$$a_{T+1/T} = T_{T+1} a_T , \tag{11}$$

and so the one step ahead prediction is given by

$$\tilde{y}_{T+1/T} = Z_{T+1} a_{T+1/T} , \tag{12}$$

Now, if we consider the problem of predicting several steps ahead, that is predicting future values for the moments $T+2, T+3, \ldots, T+h$, where h is the prediction horizon, replacing repeatedly on the transition equation, at the moment $T+h$ we have,

$$\alpha_{T+h} = (\prod_{j=1}^{h} T_{T+j})\alpha_T + \sum_{i=1}^{h-1} (\prod_{i=j+1}^{h} T_{T+i})(R_{T+j}\eta_{T+j}) + R_{T+h}\eta_{T+h},\ h = 2,3,... \tag{13}$$

On the other hand, it is known that the estimator MMSE of a_{T+h} at the moment T is the conditional expected value of a_{T+h}. Applying conditional expectation at time T to (13) we have

$$E(\alpha_{T+h} / T) = a_{T+h/T} = (\prod_{j=1}^{h} T_{T+j}) a_T . \tag{14}$$

The conditional distribution of a_{T+h} is Gaussian and its variance and covariance matrix $P_{T+h/T}$ can be obtained from (13) and (14). In the case that it does not change over time, the right expression would be

$$P_{T+h/T} = T^h P_T T^{h}{'} + \sum_{j=0}^{h-1} T^j RQR' T^{j}{'}, \quad h = 1,2,\dots \tag{15}$$

The MMSE estimator of y_{T+h} can be obtained directly from y_{T+h}, taking conditional expected values to the measure equation at time $T+h$, being

$$E(y_{T+h}/T) = \tilde{y}_{T+h/T} = Z_{T+h} a_{T+h/T}, \quad h = 1,2,\dots. \tag{16}$$

The MSE matrix is

$$\text{MSE}(\tilde{y}_{T+h/T}) = Z_{T+h} P_{T+h/T} Z_{T+h}^{'} + H_{T+h}, \quad h = 1,2,\dots. \tag{17}$$

We have to remind that the MSE matrixes $P_{T+h/T}$ do not take into account the error that comes from estimating unknown parameters from the system matrixes T_t and other parameters.

To probe if a model is right not only we have to keep in mind the goodness of fit but also it is very important to probe its goodness to predict because it is well known that not every model that fit well are also useful to predict.

Using all these residuals we calculate the CUSUM and we build the Chow test inside the sample. When we carry out a post–sample predictive test, the estimations of the explanatory variables or intervention variable coefficients are the same of those in the final state of the sample. The corresponding standardized residuals are called v_t for $t = T+1,\dots,L$. A characteristic post–sample test is given by

$$cusumt = L^{-1/2} \sum_{j=1}^{L} v_{T+j}, \tag{18}$$

which is distributed as a t_{T-L-d^*}.

Other two very useful measures to probe the predictive behaviour of the models are:

a) For the sample period, the MAPE ("Mean absolute percentage error").
b) For the post–sample period, the PSMAPE ("Post-sample absolute percentage error").

4. "SPLINE" TECHNIQUE

The original idea is owe to Schoenberg (1946), and it is useful when we have to represent non-linear behaviour in the data. In daily time series, the non-linear periodical annual behaviour requires a large number of parameters in order to be adequately represented. But structural change that takes place slowly can be represented by a regression in several steps using a cubic "spline" (Poirier, 1973). The principal advantages of this technique are:

I. It preserves the continuity of the estimated function by step functions.
II. It is very easy to use in computer systems.

4.1 "Spline" and regression models

In a non-linear regression model we suppose that,

$$y_t = f(x_t) + \varepsilon_t, \quad t = 1, \dots, T, \quad \varepsilon_t \sim \mathrm{IID}(0, \sigma^2). \tag{19}$$

Instead of adjusting $f(.)$ for only one curve, we approximate this using a cubic "spline" $g(x)$. We define k cubic functions based on $(x_0, \dots, x_k)$ (those individual points x are known as "knots") and the corresponding level values $f(x_j) = \gamma_j, j = 0, \dots, k$ in the observed points. In the case of a periodical "spline" $\gamma_0 = \gamma_k$, $d_1(x_0) = d_k(x_k)$ and $d_1(x_0^2) = d_k(x_k^2)$.

Given the "knots" and the values associated with the level, it can be shown that any point over the "spline" function" is a lineal combination of γ_j in a way,

$$g(x) = w' \gamma', \tag{20}$$

w_j is a $k \times 1$ vector which depends on the location of the "knots" and the distance between $h_j = x_j - x_{j-1}$ and the values of x_j. The algebra involving this is quite simple.

The γ_j are estimated by ordinary least squares

$$\hat{\gamma} = \sum_t (w_t \; w_t')^{-1} \sum_t (w_t \; y_t). \tag{21}$$

Therefore, an estimation of the non-linear effects is transformed into a regression problem using cubic "splines".

4.2 "Spline" technique applied to daily time series

For time series with daily observations, let ψ_t be the annual seasonal component.

To make a parameterization of ψ_t by a cubic "spline" means that we have,

1) $x_j = j, \quad j = 1, 2, \dots, s$, for the case of daily series $s = 365$ and $k < s$. In this situation we have $k = 6$, $x_0 = 1$ and $x_k = 365$.

2) $\psi_t = \psi_j$ when the *j*-th seasonal effect is present.

3) $\psi_t = \boldsymbol{w}_j' \, \boldsymbol{ø}$.

$\boldsymbol{w}_j$ is the weighting vector that depends on the "knots" and the index j and $\boldsymbol{ø}' = (\psi_1, \dots, \psi_6)$ is the parameter vector that we have to estimate.

To avoid any trouble we need to have,

$$\sum_{j=1}^{6}\psi_j = \sum_{j=1}^{6}\boldsymbol{w}_j^{'}\ \boldsymbol{ø} = \boldsymbol{w}\ \boldsymbol{ø}' = 0. \tag{22}$$

This restriction is equivalent to,

$$\psi_j^* = -\sum_{j=1}^{5}\left(w_i^*/w_k^*\right)\psi_i\,, \tag{23}$$

where

$$w^* = \sum_{i=1}^{6} w_i\,, \tag{24}$$

the seasonal component is expressed as a function of $(k-1)$ parameters.

Then we have

$$\psi_i = z_i\ \psi^*, \tag{25}$$

where

$$\psi^* = \left(\psi_1,\ldots,\psi_{k-1}\right),$$

z_j is a $(k-1)\times 1$ vector whose *i*-th element is given by

$$z_{ji} = w_{ji} - \left(w_{jk}\ w_j^*/w_k^*\right), \tag{26}$$

where w_{ji} is the *i*-th element of w_j.

In the case of the energy demand we pick x_{j-1}, $j=1,\ldots,6$:

$x_0 = 1,\quad x_1 = 71,\quad x_2 = 137,\quad x_3 = 237,\quad x_4 = 325,\quad x_5 = 365$.

Which implies

$h_1 = 70,\quad h_2 = 66,\quad h_3 = 100,\quad h_4 = 88,\quad h_5 = 40$.

To calculate the series z_{jt} with $j=1,\ldots,4$ and $t=1,\ldots,1461$, we use the IML procedure of SAS package. Once we calculate z_{jt} this will act as explanatory variables in the state space model. For the leap year 1996 we repeat twice the value of *z* for 28th February.

5. STATE SPACE MODEL FOR THE DAILY ENERGY DEMAND SERIES

The selected model turns out to be,

$$\begin{aligned} y_t &= \mu_t + \gamma_t + \sum_{i=1}^{4}\theta_{it} + \delta_1 x_{1t} + \delta_2 x_{1t}^2 + \sum_{i=1}^{4}\psi_i z_{it} + AR(1) + \chi_{1t} \\ &\quad + \upsilon_1 O_1 + \upsilon_2 O_2 + \upsilon_3 O_3 + \upsilon_4 O_4 + \upsilon_5 O_5 + \varepsilon_t\,, \qquad t = 1,\ldots,1461, \end{aligned}$$

with

$$\mu_t = \mu_{t-1} + \eta_t \quad \text{(stochastic level)},$$

$$\gamma_t = -\sum_{j=1}^{6} \gamma_{t-j} + w_t \quad \text{(weekly stochastic seasonality)},$$

and

$$\sum_{i=1}^{4} \theta_{ij} \quad \text{(fixed holiday effect)},$$

$$x_{1t} \quad \text{(explanatory variable: temperature)},$$

$$\sum_{i=1}^{4} \psi_{ij}\ z_{ij} \quad \text{(“spline” of fixed seasonality)},$$

$$\begin{bmatrix} \chi_{1t} \\ \chi^*_{1t} \end{bmatrix} = \rho_1 \begin{bmatrix} \cos\lambda_c & \operatorname{sen}\lambda_c \\ -\operatorname{sen}\lambda_c & \cos\lambda_c \end{bmatrix} \begin{bmatrix} \chi_{1,(t-1)} \\ \chi^*_{1,(t-1)} \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa^*_t \end{bmatrix}, \quad \text{(stochastic cycle)},$$

$$AR(1) = \begin{bmatrix} \chi_{2t} \\ \chi^*_{2t} \end{bmatrix} = \rho_2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \chi_{2,(t-1)} \\ \chi^*_{2,(t-1)} \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa^*_t \end{bmatrix}, \quad \text{(random } AR(1) \text{ cycle)},$$

$$O_1 = \begin{cases} 1 & \text{if } t = 8/12/95 \\ 0 & \text{otherwise} \end{cases}, \quad O_2 = \begin{cases} 1 & \text{if } t = 28/9/96 \\ 0 & \text{otherwise} \end{cases},$$

$$O_3 = \begin{cases} 1 & \text{if } t = 24/12/96 \\ 0 & \text{otherwise} \end{cases}, \quad O_4 = \begin{cases} 1 & \text{if } t = 15/8/97 \\ 0 & \text{otherwise} \end{cases},$$

$$O_5 = \begin{cases} 1 & \text{if } t = 31/12/98 \\ 0 & \text{otherwise} \end{cases}, \qquad (27)$$

$O_i,\ i = 1,\ldots,5,$ are the ”outliers”,

$\varepsilon_t \sim NID(0, \sigma^2_\varepsilon)$, $\eta_t \sim NID(0, \sigma^2_\eta)$, $w_t \sim NID(0, \sigma^2_w)$, κ_t and $\kappa^*_t \sim NID(0, \sigma^2_\kappa)$ are the random effects and are mutually independent.

Table 1: Model components estimation of final state vector at the end of the period

Coefficient	Estimate	*t*-value	*p*-value
μ_t	225.21	66.606	0.0000
χ_{1t}	-2.2796		
χ_{1t}^*	-0.1400		
γ_{1t}	-11.247	-19.987	0.0000
γ_{2t}	9.8745	17.703	0.0000
γ_{3t}	9.1415	16.406	0.0000
γ_{4t}	9.1976	16.549	0.0000
γ_{5t}	9.3016	16.727	0.0000
γ_{6t}	5.8841	10.52	0.0000
AR(1)	-1.9964		
θ_1	-32.403	-46.594	0.0000
θ_2	-35.040	-48.994	0.0000
θ_3	-12.451	-7.3214	0.0000
θ_4	-5.9580	-3.9769	0.0001
δ_1	-2.9310	-17.591	0.0000
δ_2	0.0960	22.318	0.0000
ψ_1	-6.6402	-3.96	0.0001
ψ_2	-2.8927	-1.3919	0.1642
ψ_3	-6.3385	-3.7472	0.0002
ψ_4	9.7487	6.2248	0.0000
Outlier 8/12/95	-25.389	-8.5888	0.0000
Outlier 28/9/96	12.165	4.1352	0.0000
Outlier 24/12/96	-12.852	-4.3021	0.0000
Outlier 15/8/97	8.0946	2.7514	0.0060
Outlier 31/12/98	-12.677	-4.2728	0.0000

This model indicates that there is a random level. The weekly stochastic seasonality is represented by “dummy” variables and the annual seasonality is represented by “spline” functions. The temperature appears in a linear as well as in a quadratic way, the holiday effect is quite important. We have two cycles; one of them is autoregressive and there are five “outliers”. The estimation of the model can be seen in the Table 1. All the coefficients are significant, except those corresponding to the second “spline”.

The goodness of fit measures are shown below in Table 2 being all of them adequate.

Table 2: Goodness of fit measures of the model

H	$\tilde{s}$	C	DW	Q(37,30)	R_s	AIC
1.071	3.885	1.16	1.944	42.24	0.8271	2.7539

The estimation of the seasonal coefficients at the end of the period are highly significant. These can be seen on the following Table 3.

Table 3: **Estimation of the weekly seasonal coefficients at the end of the period**

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
$\hat{g}$	5.884	9.302	9.198	9.142	9.874	-11.250	-32.150

There is a different behaviour on Mondays from the rest of the days of the week, and there is also a rapid descend of the demand of energy during the weekend.

The Figure 4 shows the estimated trend of the model.

5.1 Predictive goodness of the model

It is as important to verify the predictive goodness of the model as it is to probe its goodness of fit. Therefore we present the following tests performed on the model.

Firstly we made prediction tests inside the sample. The CUSUM graphic (Figure 5) shows that the values for the last year of observation are within the confidence bands. Chow test and *t* test are not significant (see Table 4), in consequence the model overcomes these predictive tests.

Figure 4**: Estimated trend of the model**

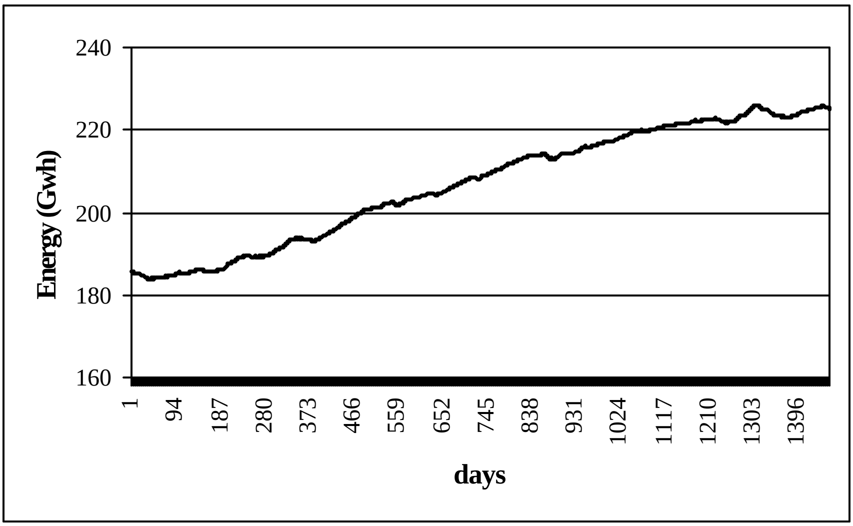


Table 4: Predictive tests within sample

Period	Chow *F* test		*t*-test	
	d.f.	*F*-value	d.f.	*t*-value
01-08-98 to 31-07-99 (1 year)	(364, 1089)	1.166	(1089)	0.822
01-02-99 to 31-07-99 (6 months)	(180, 1274)	1.022	(1274)	0.604
01-05-99 to 31-07-99 (3 months)	(90, 1364)	0.810	(1364)	0.819
01-07-99 to 31-07-99 (1 month)	(30, 1424)	1.337	(1424)	0.1645
25-07-99 to 31-07-99 (1 week)	(7, 1447)	1.242	(1447)	-1.4937

The values of the MAPE coefficients are less than 2 (see Table 5), therefore we can say that the average forecast error, made one step ahead, is less than 2% of the predictive value in all the periods under study.

Table 5: MAPE coefficients of the model

Period	Coefficient
01-08-98 to 31-07-99 (1 year)	1.54
01-02-99 to 31-07-99 (6 months)	1.46
01-05-99 to 31-07-99 (3 months)	1.35
01-07-99 to 31-07-99 (1 month)	1.68
16-07-99 to 31-07-99 (15 days)	1.54
25-07-99 to 31-07-99 (1 week)	1.87

The post-sample forecast with its confidence interval for the horizon $h = 1$ to $h = 43$ are represented in the Figure 6 below.

Figure 5: CUSUM for the last year of observation

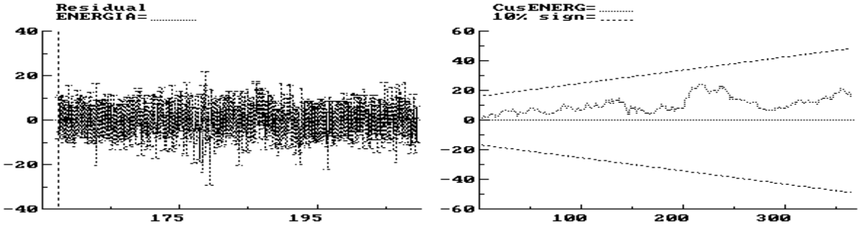
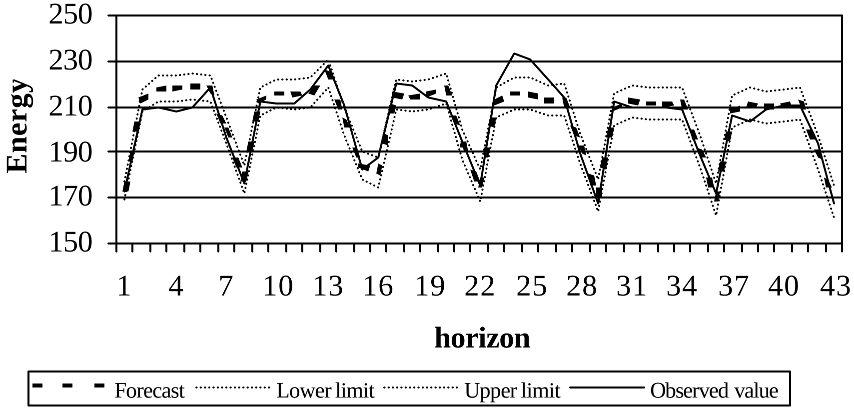


Figure 6: Post-sample forecast and confidence bands



The values of the PSMAPE coefficient are less than 2.5 (see Table 6), this indicates that the forecast errors in the post-sample period, up to 43 steps ahead do not overcome the 2.5% of the predicted value. The real values are mostly in the confidence interval, and only for horizons 22 to 24 are above the upper limit.

Table 6: PSMAPR coefficients

Period	Coefficient
01-08-98 to 31-07-99 (1 year)	1.54
01-02-99 to 31-07-99 (6 months)	1.46
01-05-99 to 31-07-99 (3 months)	1.35
01-07-99 to 31-07-99 (1 month)	1.68
16-07-99 to 31-07-99 (15 days)	1.54
25-07-99 to 31-07-99 (1 week)	1.87

All the results found show that the model can be considered as highly satisfactory to make predictions.

Despite that, we must realise that the entire forecast are made knowing the future values of the temperature variable. This, in general, is not fulfilled in real life because when we make that prediction those values are not available.

5.2. Alternative model when the temperature is unknown

It can be harder to predict future temperature values than predict the energy demand series. To solve this problem we propose the following two solutions:

I. Make predictions for different settings according to possible temperature values, because the rank of those possible values is pretty narrow in the short run.
II. Propose forecasts for different settings knowing that only extreme temperatures affect the energy demand. Therefore we can introduce “dummy” variables that reflect this kind of behaviour. Since the forecasts for daily series are only for the short run, it is easier to determine if the temperature will be extreme in the following days.

In our case, we use a model with three “dummy” variables; one for temperatures above 28º C, one for temperatures below 8º C and one for standard temperatures. We incorporate them to the model instead of the polynomial of second degree we are using for temperature.

The only crucial and significant variable is the one that reflect temperatures above 28º C. Despite the fact that the goodness of fit statistics are lower than the corresponding to the previous model, we have to say that the trend has not periodical components. The predictive goodness inside the sample is very good, the CUSUM for the last year of observations is inside the confidence interval, the MAPE coefficient is 1,63 and Chow test is unimportant. In consequence, we can make use of this model when we do not know the exact values of the temperature.

6. CONCLUSIONS

This work shows the usefulness of space state models in fitting and predicting daily time series, as well as the convenience of using the “spline” technique for the treatment of annual seasonality.

The proposed model happens to be highly satisfactory and its most important features are:

1) The fitting explains more than 80% of the series variability.
2) The “spline” technique allows us to find the behaviour of the trend component free of any periodical component.
3) The forecast errors inside the sample (one step ahead) for different values are less than 2% of the original ones.
4) The post-sample forecast errors made up to 43 periods ahead are less than 2,5% of the original values.
5) The model allows us to interpret in a simple fashion the weekly seasonal effect that shows the different behaviour of Mondays and weekends.
6) The holiday effect changes according to the day of the week the holiday occurs.

7) The relation with temperature shows its non-linear effect, increasing the energy demand when we have extreme temperatures, and even more when the temperatures are high.

We also recommend alternative methods when future temperature values are unknown. In the literature there are models to fit and forecast daily series; among them we have ARIMA models or regression models. We can use them for future studies. Another line of research would be to find models for other frequencies, and compare their behaviour. We can try models for different regions, because it is likely that they have different behaviours.

7. ACKNOWLEDGEMENTS

We are very grateful to María de la Mercedes Abril for her helpful collaboration in the preparation of this paper. I am also thankful to the editor for his healthy comments.

REFERENCES

1. Abril, Juan Carlos. (1999), *Análisis de Series de Tiempo Basado en Modelos de Espacio de Estado*, Buenos Aires: EUDEBA.

2. Box, G. E. P. and Jenkins G. M. (1970), "*Times Series Analysis Forecasting and Control*", San Francisco: Holden Day.

3. Gordon, F.(1996). *Previsao de Carga Diária Através de Modelos Estruturais Usando Spline*", Ph. D. Thesis. Pontifícia Universidade Católica de Rio de Janeiro, Rio de Janeiro.

4. Harvey, A. C. (1989). *Forecasting Structural Time Series and the Kalman Filter*". Cambridge (U. K.): Cambridge University Press.

5. Harvey, A. C. and Koopman, S. J. (1993), "Forecasting Hourly Electricity Demand Using Time Varying Splines. *Journal of the American Statistical Association*. **88**, 1228-1236.

6. Harvey, A. C., Koopman, S. J. and Riani, M. (1996), "The Modelling and Seasonal Adjustment of Weekly Observations", *Journal of Business and Economic Statistics*, **15**, 354-368.

7. Koopman, S. J., Harvey, A. C., Doornik, J. A. and Shephard, N. (1995), *STAMP 5.0: Structural Time Series Analyser, Modeller and Predictor*, London: Chapman and Hall.

8. Poirier, D. J. (1973), "Piecewise Regression Using Cubic Spline", *Journal of the American Statistical Association*, **68**, 515-524.

9. Schoenberg, I. J., (1946), "On the Problem of Smoothing or Graduation. A First Class of Analytic Approximation Formulae", *Quarterly Journal of Applied Mathematics*, 45-49.