versão impressa ISSN 0101-7438 / versão online ISSN 1678-5142

A METHODOLOGY TO FILTER TIME SERIES: APPLICATION TO MINUTE-BY-MINUTE ELECTRIC LOAD SERIES

Mayte Suarez-Farinas
Rodrigo Lage de Sousa
Reinaldo Castro Souza *
Electric Engineering Department
Pontificia Univ. Católica do Rio de Janeiro (PUC-Rio)
Rio de Janeiro – RJ
reinaldo@ele.puc-rio.br

* *Corresponding author* /autor para quem as correspondências devem ser encaminhadas

Recebido em 11/2002; aceito em 08/2004 após 1 revisão
Received November 2002; accepted August 2004 after one revision

Abstract

In this article a methodology for filtering a time series is presented, with application to high frequency series such as the minute-by-minute electric load series. The goal of this approach is to detect and substitute the irregularities of the time series that can produce distortions on the modelling stage. Outlier values are detected through a dynamic linear model and the Bayes factor tool; missing values are then interpolated with a Smoothing Cubic Spline. The performance of the proposed approach is illustrated using real data and evaluated through a series of tests where the irregularities have been simulated.

Keywords: load series; DLM model; irregularities in time series; cubic spline; Bayes factor.

Resumo

Neste artigo apresenta-se uma metodologia para a filtragem de séries temporais, com aplicação em séries de alta freqüência. Esta metodologia tem como objetivo detectar e substituir as irregularidades da série temporal que podem comprometer a etapa de modelagem. São detalhados o modelo linear dinâmico utilizado para detectar os valores *outliers* e o emprego do Fator de Bayes. Na interpolação de valores faltantes utiliza-se o *Spline* Cúbico Suavizado. O desempenho da metodologia proposta é avaliado a través de vários testes onde as irregularidade foram simuladas.

Palavras-chave: série de cargas; modelo linear dinâmico; irregularidades em séries temporais; *spline* cúbico; fator de Bayes.

1. Introduction

Following a worldwide trend, the Brazilian electric sector has been undergoing intense transformation over the last few years. In this new context, load forecasting for very short periods assumes vital importance, serving as a base not only for the calculation of future pricing of electric power, but also for the programming of optimum dispatch of the plants in the system carried out by the National Operator of the Electric Power System (ONS). Besides this operational use of these forecasting, they are also important in the routine short-term transactions of the pool electricity market, which, in Brazil, is under responsibility of another private body known as MAE (Wholesale Electricity Market).

Aimed at developing a profile for a computational system to forecast short-term electric load series for the ONS, the Cahora Project was created in a partnership between Cepel and PUC-Rio. This system and the implemented models are detailed in Rizzo (2001). The first great challenge in developing forecasting models was the great amount of irregularities found in the data retrieval system, described as follows.

CNOS, the ONS headquarters located in Brasilia, is responsible for the consolidation of load measurements for electric energy utilities in the Brazilian system. The readings of the measurement points arrive at the data retrieval centre by telemetry, at approximately 20-second intervals. Every minute, a recording is made of the available value in the reading. This system of data retrieval is subject to the occurrence of irregularities that can be motivated by erroneous information transmission, errors in data consolidation or unexpected load behaviour (blackout). These irregularities can mainly be described as:

- Occurrence of missing values: when there is no load-value readings during the interval of one minute.
- Occurrence of outliers: when the registered value is nonsense, outside the range of expected behaviour. This can be due, for example, to measurement failure at some point of the interchange or inherent problems in the information sending process. This category also includes real values corresponding to unexpected events, such as load failures or blackouts.

These irregularities can impair the performance of any forecasting model. Thus, before carrying through the half-hour load series modelling, a data treatment module has been developed. Using a filter for Missing Values and Outliers (FMO) these irregularities are corrected, transforming the minute by minute data recording into a statistically modelled time series, that can be aggregated to create higher periods series (for example, at half-hour intervals).

The FMO has two main functions: that of detection of Outliers, declaring them to be missing values; and the interpolation of all resultant missing values. Once the values have been declared to be missing, these values must be substituted by values in conformity to the dynamics of the series, to pass on to the modelling stage. Since the system works as a daily basis operation, it will filter out the data obtained throughout the day on a minute-by-minute basis, at the end of the day.

After the application of the FMO to the minute-by-minute series, some irregularities still persist in the filtered series, for example abrupt reductions of the series variance. These irregularities are outside the expected pattern. Therefore, a procedure called the “Pattern Filter” (PF) was implemented to detect and substitute these irregularities. The PF is based on matching up the daily load curves with standard patterns established by neural network techniques.

In this article we describe the methodology implemented for outlier detection and missing value interpolation. Theoretical aspects related to the Dynamic Linear Models and Bayes factor, both used in outlier detection procedure and Smoothing Cubic Spline, commonly used to interpolation, are discussed in section 2 and 3. A Pattern Filter is presented in section 4 and the whole procedure is illustrated in section 5 using real electric load data. Finally, results obtained in the simulations carried out to evaluate the performance of the proposed methodology are presented in Section 6.

2. Detection of Irregularities

To model the dynamics of the time series, we used a Bayesian Linear Model with a discounting factor for the trend and the slope and a scale factor for the variance of the observations. Canton (1999) used a simpler model with a similar purpose but the trend components were not considered. For outlier detection, we used the Bayes factor, as described in West & Harrison (1986). Because some important discontinuities did not fit into this first implementation, such as the irregularities that will be described further ahead, some modifications to the original procedure have been made in order to detect them. Next, we briefly describe the outlier detection process, including a brief summary about the dynamic linear model setup and Kalman Filter equations.

2.1 The Dynamic Linear Model (DLM)

Given a time series Y_t, the DLM is described by:

$$\begin{aligned} Y_t &= F_t'.\theta_t + \nu_t \qquad \nu_t \sim N[0, V_t] \\ \theta_t &= G_t.\theta_{t-1} + \omega_t \quad \omega_t \sim N[0, W_t] \end{aligned} \qquad \text{corr}(\omega_t,\nu_t)=0 \qquad (2.1)$$

where time t=1...,T; Y_t is the vector of observations, with dimension nx1; θ_t is the vector of parameters with dimension qx1 containing the non-observable components of the model (level, slope, seasonal component, etc.); ω_t and ν_t are independent and normally distributed random errors, uncorrelated to one another.

In the Bayesian formulation of this model the variances of the errors, given by the matrices V_t and W_t, must be specified. The problem of evolution of error variances must be solved, when we assume that they are unknown. Pole *et al.* (1994) and West & Harrison (1997) proposed a learning procedure for the unknown observation variances V_t, in terms of a scale factor ϕ that, in the case of normality, is the observation precision. The idea is to scale all variances and covariances in the updating equations of the DLM using ϕ. The recursive equations of the Kalman Filter are still valid; however, conditional distributions of ϕ will be normal while the marginal distributions will be t-Student with the degrees of freedom given by the corresponding marginal distribution of ϕ. Working with $V_t=\phi^{-1}$, the model (2.1) can be rewritten as:

$$\begin{aligned} Y_t &= F_t'.\theta_t + \nu_t \qquad \nu_t \sim N[0, \phi^{-1}] \\ \theta_t &= G_t.\theta_{t-1} + \omega_t \quad \omega_t \sim N[0, W_t^*\phi^{-1}] \end{aligned} \qquad (2.2)$$

We determine the matrix W_t through the formulation of discounting factors, introduced by Harrison (1965). The idea is based on the consideration that the informative content of one

observation decays with age. Thus, the prior variance for time instant t is calculated as a function of the posterior variance in time t-1, determined by a discount factor δ_0 ($0 < \delta_0 \leq 1$). The discount factor represents the amount of information lost with time as the series evolves. This is equivalent to establishing the variance of errors of the state vector as $W_t=(\delta_0^{-1}-1)G_tC_{t-1}G_t$, where C_{t-1} is the variance of the posterior state vector. In this case, different discount factors for each component of the state vector are considered, as in Ameen & Harrison (1985). Thus, the vector of discount factor will be $\delta=(\delta_1,\delta_2,\ldots,\delta_n)$.

For a 1-day long minute-to-minute load series, we consider a local linear trend model (F=[1,0], G=[1 1;0 1] and values of 0.9 and 0.8 were used for discount factors of level and trend respectively. So, 10% and 20% of level and trend information decays during each interval of time. (See Ameen & Harrison (1985) for comments about discount factor selection).

2.2 Kalman Filter

In what follows, we summarize the recursive equations of the Kalman Filter for this model. The distributions conditioned to values of ϕ are the same as in the DLM with known variances and the marginal distributions are obtained from the former. The interested reader may find details about the dynamical model elsewhere, in particular in Pole *et al*. (1994) and Harrison & West (1997).

Initialization

At t=0, we assume $(\theta_0|D_0,\phi)\sim N\left[m_0,C_0^*\phi^{-1}\right]$, $(\phi|D_0)\sim G\left[n_0/2,d_0/2\right]$

Priors on time t

State Vector: $(\theta_t \mid D_{t-1},\phi) \sim N[a_t, R_t^*\phi^{-1}]$, where $R_t*\phi^{-1} = R_t$.

Scale Factor: $(\phi_t \mid D_{t-1}) \sim G[n_{t-1}/2, d_{t-1}/2]$, with mean $1/S_{t-1}$, where $S_{t-1}=d_{t-1}/n_{t-1}$.

The unconditional (on ϕ) prior for state vector: $(\theta_t \mid D_{t-1}) \sim T_{n_{t-1}}[a_t, R_t]$, where $R_t=R_t*S_{t-1}$. Here, a_t and R_t are calculated as in the known variance model, considering W_t calculated through the discount factors. Thus: $a_t = G_t m_{t-1}$, $R_t = \delta^* G_t C_{t-1} G_t' \delta^*$ where $\delta^* = \text{diag}(\delta_1^{-1/2}, \delta_2^{-1/2}, ..., \delta_n^{-1/2})$ is the inverse of the discount matrix.

One-step-ahead forecasting

Conditional distribution: $(Y_t|D_{t-1},\phi) \sim N[f_t, Q_t^*\phi^{-1}]$ with $f_t = F_t'a_t$, $Q_t^* = 1 + F_t'R_t^*F_t$.

The unconditional distribution (on ϕ): $(Y_t|D_{t-1}) \sim T_{n_{t-1}}[f_t, Q_t]$ with $Q_t=S_{t-1}Q^*_{t-1}$.

This result is similar to the one obtained when a known variance was considered, if we consider the expected value of prior S_{t-1} as a point estimation for ϕ until time t-1:

$$Q_t = S_{t-1} + F_t'R_t^*S_{t-1}F_t \quad \text{or} \quad Q_t = S_{t-1} + F_t'R_tF_t$$

Posterior (Update)

Scale Factor: $p(\phi|D_{t-1},Y_t) \sim G[n_t/2,d_t/2]$, where $n_t = n_{t-1}+1$, $d_t = d_{t-1}+e_t^2/Q_t^*$, e_t innovations.

State Vector: $(\theta_t|D_t,\phi) \sim N[m_t, C_t^*\phi^{-1}]$, where $m_t = a_t + R_t^* F_t e_t/Q_t^*$

$$C_t^*\phi^{-1} = \phi^{-1}\left(R_t^* - R_t^* F_t F'_t R_t^*/Q_t^*\right)$$

The unconditional distribution (marginal): $(\theta_t|D_t) \sim T_{n_t}[m_t, C_t]$, where $C_t = C_t^* S_t$ and S_t is the point estimation for the scale factor. Under these considerations, and taking into account that $R_t = R_t^* S_{t-1}$, the equations can be rewritten as:

$$m_t = a_t + R_t F_t e_t/Q_t, \quad C_t = \frac{S_t}{S_{t-1}}\left(R_t - R_t F_t F'_t R_t^{-1}/Q_t\right)$$

2.3 Bayes Factor

The Bayes factor concept was introduced by Jeffreys (1961), and constitutes a very useful tool for testing hypotheses and model selection. The Bayes factor is none other than the ratio between the likelihood function of the standard model (Model **S**) and the likelihood of an alternative model (Model **A**). In the time series context it is used as a measure of the predictive capacity of the model. A thorough discussion, in the classical and bayesian context with multiple references and applications can be found in Kass & Raftery (1995).

West (1986) considers a procedure for a sequential monitoring of the model, based on the Bayes factor. This procedure, also discussed in West & Harrison (1986), continuously validates the predictive capacity of the model, under the bayesian approach. Since the objective is to develop a tool sensitive to local failure in the model, the Bayes factor concept was adapted.

Let A be the alternative model and $p_A(Y_t|D_{t-1})$ be its predictive distribution. Then we can define:

$$\Omega_t(k) = p(y_t,...,y_{t-k+1}|D_{t-k}) / p_A(y_t,...,y_{t-k+1}|D_{t-k}) \qquad 1 \le k \le t \qquad (3.1)$$

If we consider $\Omega_t(0)=1\ \forall\ t$, we obtain a recursive equation for $\Omega_t(k)$: $\Omega_t(k) = H_t.\Omega_{t-1}(k-1)$, where H_t is the Bayes factor based only on y_t: $H_t = p(y_t|D_{t-1}) / p_A(y_t|D_{t-1})$.

In our case, the predictive distributions are t-Student $(Y_t \mid D_{t-1}) \sim T_{n_{t-1}}[f_t, Q_t]$. If we consider the alternative model, based on the use of "power discounting" as in Smith (1979), we will assume a diffuse predictive density, that is, with the same mean and variance Qa_t such that $Q_t = \rho Qa_t$. Then, the Bayes factor becomes:

$$H_t = \frac{\left[n_{t-1} + \frac{(Y_t - f_t)^2}{Q_t}\right]^{-\frac{n_{t-1}+1}{2}}}{\left[n_{t-1} + \frac{(Y_t - f_t)^2}{Qa_t}\right]^{-\frac{n_{t-1}+1}{2}}} = \left[\frac{n_{t-1} + e_t}{n_{t-1} + \rho e_t^2}\right]^{-\frac{n_{t-1}+1}{2}} \qquad (3.2)$$

Given an adjusted alternative, a value less than 1 and significantly small for H_t is an indication of the existence of discontinuity in the series. H_t indicates the existence of a potential outlier, but in cases of small or gradual changes, individual Bayes factors (based only on y_t) may not be small enough to indicate failure in the model. Thus it is necessary to use jointly $\Omega_t(k)$, whose small values suggest a possible change in the past. The algorithm proposed by West (1986) is summarized as follows: At time t, whatever may have previously occurred, proceed as follows:

Step 0. Select values of $\tau > 0$, $\rho \in [0,1]$ and $Q \in IN$.

Step 1. Compute in sequence quantities L_t and ℓ_t defined by:

$L_t = H_t \min(1, L_{t-1})$ and ℓ_t, the run-length parameter, as $\ell_t = \begin{cases} \ell_{t-1} & if \quad L_{t-1} < 1 \\ 1 & if \quad L_{t-1} \geq 1 \end{cases}$

Step 2. Determine irregularities

A. Assessment of S_t

If $H_t \geq \tau$, y_t is consistent with the standard model S_t, go to B.

If not, ($H_t < \tau$), y_t is potentially an outlier; go to 3A.

B. Assessment of S_t, S_{t-1}, …

If $L_t \geq \tau$, the sequence of standard model is satisfactory (system controlled). Go to 3B.

Or else ($L_t < \tau$,) or ($\ell_t > Q$); then, we have structural breaks. Go to 3A.

Step 3. Update

A. Reject observation y_t. Update Kalman Filter equation allowing for change: incrementing the uncertainty of state vector: $m_t = a_t$, $C_t = \alpha C_t$

B. The sequence of standard model is satisfactory. Standard update of Kalman Filter equation.

When $\ell_t = 1$ the discontinuity is of the "transient" type. Value $\ell_t > 1$ indicates a structural change that started in instant $t - \ell_t + 1$. When $\ell_t > Q$, even when $\ell_t > Q$, we can admit that at instant t, a slow structural change is occurring, from instant $t - \ell_t + 1$.

It is still necessary to establish the values for τ (real positive), ρ (real between 0 and 1) and Q (natural). A discussion of this choice can be found in West & Harrison (1986) and West (1986). Following them we will be working with $\rho = 0.15$ (which implies that the standard deviation of model A is roughly 2.5 times the standard deviation of model S) and $\tau = .2$, which indicates that for standardized prediction residuals around 2.5, $H_t = \tau$ and Y_t alone provides enough evidence to reject S_t. A parameter Q controls the threshold to declare structural breaks. Considering the high frequency of the data, we worked with $Q = 6$. Values of Q smaller than 6 will cause frequent declarations of structural breaks, when in fact there are several consecutives outliers. Finally α, which represents the uncertainty about the state vector was set as 1.5 in step 3-A.

Considering the characteristics of the short-term load series, certain alterations to West's algorithm had to be considered. In West's algorithm, when we are in the presence of any type of discontinuity the update of the model must be made in such a mode as to allow for changes in the model. The variance matrix of posterior density is multiplied by a factor greater than one. This factor can be a matrix, allowing a different inflation for the posterior variance of each component. This is also advisable in the case of missing values.

This algorithm should identify outliers and structural changes. In our case, the minute-by-minute load series frequently has 2 and up to 4 consecutive outliers, without this implying a structural change. An adaptation of the model, that increased the uncertainty of the model when faced with an outlier, would make it difficult to detect other outliers later on. As our goal is not modeling but the detection of discontinuities, we propose to increase the uncertainty only when a structural change is detected, keeping the uncertainty unchanged when an outlier is detected. Furthermore, our data presents large intervals with consecutive missing values; consecutive increments to the variance matrix would make it diverge numerically, making the system unavailable for outlies detection. Thus, we decided not to increment variance in the case of missing values.

According to the nature of the data, abrupt structural changes cannot happen because the load cannot have a very strong variation from one minute to the next. Hence, both the outlier values and structural changes must be substituted. When a structural change is detected, if it is a permanent one, all the values in its neighbourhood are substituted in order to smooth the structural change.

3. Missing Values Interpolation

Some methodologies for missing values estimation in time series have been proposed the literature. Parametric alternatives have been widely used (Brubacher, S.R. *et al.*, 1976), by means of a statistical model to adjust the data, and using it later in the interpolation. Ferreiro (1987), Harvey & Pierse (1984) and Ljung (1989) use this approach as well.

A natural idea to substitute the missing values and outliers would be to use the bayesian model adjusted in the outlier declaration stage and then to substitute the extraneous observations with the one-step-ahead forecasts made in the previous instant. However, the existence of data blocks with many consecutive missing observations (or declared outliers), sometimes for more than a one hour period (60 observations), made it necessary to abandon these techniques because, as the forecast horizon grows the forecasts cease to be reliable.

Thus, we decided to use the smoothing cubic spline, previously applied to the interpolation of time series by Gordon (1996) and to the handling of missing values in Koopman (1991) and Koopman *et al.* (1998).

3.1 Smoothing Cubic Spline (SCS)

The Smoothing Cubic Spline (SCS) is a non-parametric techniques that has been applied in interpolation problems in time series (see Ferreiro, 1987; and Koopman *et al.*, 1998) with good results. This method has been used in some packages such as the SsfPack (Koopman *et al*., 1998b). However, the SCS is not well-behaved for series with a considerable number of consecutive missing values. This result is basically due to the fact that SCS generates

solutions with accented curvature when many consecutive missing values are presented, distancing itself from the pattern of the data, as we will see further ahead.

The underlying idea in cubic spline is the smoothing of a third-order polynomial. This smoothing is effected through a model in state space form that represents the spline, afterwards using the Smoothing Kalman Filter (SKF), as described in Kohn & Ansley (1987). This type of model takes into consideration all the available values and not only three points.

Weinert *et al.* (1980) defines in a general way the concept of smoothing spline, discussing the problem and establishing the existence and uniqueness of the solutions. Kohn & Ansley (1987) discuss the particular case of polynomial splines defined as follows:

Let t_1, t_2... t_N be the available time periods. The problem of interpolation by smoothing polynomial splines consists of the searching of a polynomial f that minimizes the loss function:

$$\sum_{i=1}^{N}\left\{Y_{t_i} - f(t_i)\right\}^2 + \lambda \int_0^t \{Lf(t)\}^2 dt \qquad (3.3)$$

where L is the differential operator $L=d^m/dt^m$. When m=2, we will have the smoothing cubic spline and the loss function consists of the integral of the square curvature, and the solution is such that $Y_{ti}=f(t_i)$.

Wahba (1978) and Weinert *et al.* (1980) represent the spline as the limit of the conditional expectancies of a random process. This allows Kohn & Ansley (1987), Koopman (1991) to establish efficient algorithms to adjust the smoothing cubic spline, through the recursive equations of the SKF to state space form of the random model. The reader may find a synthetic description in Durbin & Koopman (2001).

In the case where the number of consecutive missing values is not so high, spline cubic offers a good solution. But, on the other hand, when there are many consecutive missing values to interpolate, the spline does not reflect all the details of the load curve due to the softness of the solution. For this reason we propose to substitute missing values with a convex combination of cubic spline and a value which reflects the load-curve pattern.

Let Y_t be the minute-by-minute load series. Assume that Y_t is a missing value at time t=to. This missing value will be substituted by the convex linear combination of the values SY_t and TY_t. That is: $Y_{to}=\alpha TY_{to}+(1-\alpha)SY_{to}$, where S_t is the value interpolated by smoothing cubic spline and TY_t is the value of the load corresponding to the same time instant of the day with the load pattern nearest to the current day pattern. Given the load curve, the procedure to find a similar stretch to interval [t-stretch, T-1] where t is the corresponding time instant and Y_t is the value to substitute. The criterion used to select similar stretch could be the minimum mean square error of the average quadratic error, after adjusting both intervals to the same mean.

4. Pattern Filter

After the application of the Missing and Outlier Filter to the minute-by-minute series, certain irregularities will still persist in the filtered series. The load curves for large cities may show large instantaneous fluctuations, but nevertheless display daily patterns that are highly

stereotyped and reproducible. The telemetry system that feeds data to CNOS may have technical or transmission problems resulting, rarely, in a large (days) batch of data which fail to conform to this stereotyped daily pattern. It has happened, for example, that the load curve of a certain day failed to match the expected pattern for that day, without any visible outliers that could cause such a deviation. To solve this problem it was decided to implement the Filter of Pattern (FP) on the aggregated series every 30 minutes.

Pattern Filter Methodology

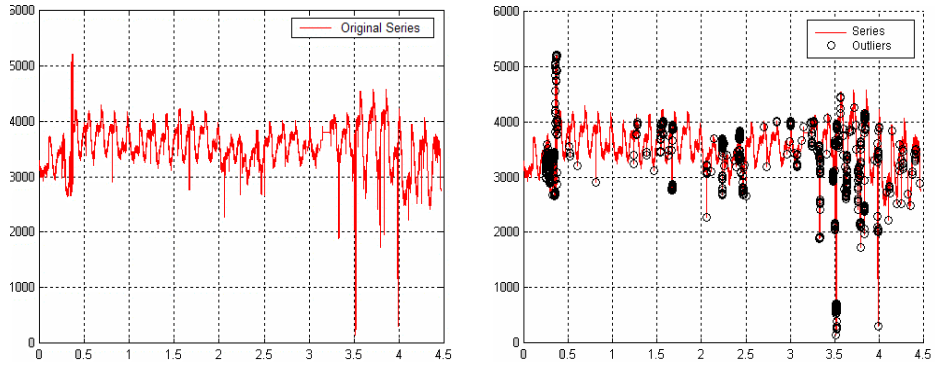
In previous work, Sobral (1999) established a classification of hourly load curves in several types for one of the Brazilian utilities using a 4x4 Kohonen Neural Network. In this work, the days were grouped according to the hourly load pattern and the prototypes were obtained. These prototypes will serve as a base for the implementation of the Filter of Pattern.

As the idea is to compare the prototypes with the load series at 30-minute intervals, the same groups have been used. From the hourly-load profiles of each group, we have obtained through linear interpolation the profiles of the load curve at each 30-minute interval. For day d we have a 30-minute load curve and then we analyze whether it follows the expected pattern. In order to do this, we select the most adequate profile using certain criteria. We use for this selection the nearest prototype to a Mean Absolute Percentage Error (MAPE). Taking into consideration that the goal is the daily pattern and not the mean of the load, we make sure that the two curves (the daily curve and the nearest prototype) have the same mean, for the MAPE computation. If the MAPE of the nearest profile is bigger than the specified threshold, we consider this day to be outside the expected standard and substitute it.

We do not propose that the whole day should be substituted because it might be the case that not all intervals of the day are responsible for the pattern failure. Thus, the day is divided into the following intervals: 00:00h to 05:30h; 06:00h to 11:30h; 12:00h to 14:30h; 17:00h to 19:30h; 20:00h to 23:30h. With them, we detect those with a MAPE bigger than the threshold. These intervals are substituted by the corresponding intervals of the profile curve that had been adjusted to have the same mean as the curve of day under study. The threshold chosen for the MAPE was 5%. At first glance, this value may seem small but we should remember that both curves have the same mean. Furthermore, to choose this value, we observed the behaviour of this statistics for a number of days, including some days that had been considered out-of-pattern by the experts.

5. Application to Real Data

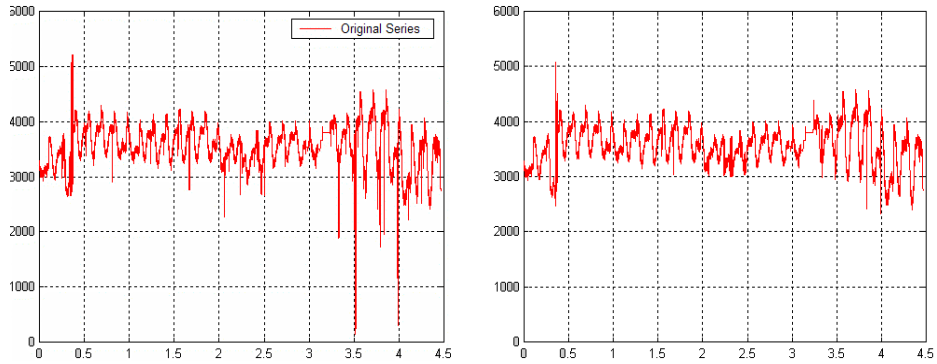
To show the behaviour of the proposed methodology for outlier detection and missing value interpolation, we present the results when applied to a minute-by-minute load series for a Brazilian Southeast utility. Due to the high volume of data we only present herein the filtering of series corresponding to August 1999. Figure 5.1 (a) shows the 44640 observations of this series, where 189 observations (0.42%) were missing. Using the outlier detection procedure discussed in section 2, 159 observations (0.34 %) were declared outliers as Figure 5.1 (b) shows.



(a) (b)

Figure 5.1 – Minute-by-minute load series. Month: August
(a) Original Series. (b) *Outliers* detected.

We then used the missing values interpolation procedure described in section 3 to substitute 348 observations (0.78%) declared missing values and outliers of the load series. The resulting series is shown in Figure 5.2 (b).



(a) (b)

Figure 5.2 – minute-by-minute load series. Month: August
(a) Original Series. (b) Series having substituted outliers and missing values.

With regard to missing values interpolation through SCS, we should warn that if the interval of consecutive missing values increases in length (say for example, by more than 100 consecutive missing values), the performance of SCS may be jeopardized. This happens because the solution offered by SCS could develop wild bends when attempting to interpolate into a large interval of missing values, and then the interpolated curve would be far from the expected curve. This problem could be circumvented by increasing the smoothing parameter λ in 3.3, yielding a smoother solution; however this would be counterproductive because it would lead us to solutions not close enough to the Y_t's, and would affect the performance of the method in most general cases.

A typical case where the Pattern Filter is required is illustrated in Figure 5.3. Here we show the series for September 1999. After the missing values and outlier filter, the series was aggregated in a 30-to-30 minutes series as shown in Figure 5.3 (a). An anomaly is obvious in the period between the 13th and 20th of September (around observations 600-1000), as an abrupt decrease in the daily variance. Power engineers assure us that this decrease in the variance did not, in fact, happen. Figure 5.3 (b) shows the resultant series after pattern filtering, showing a more plausible pattern.

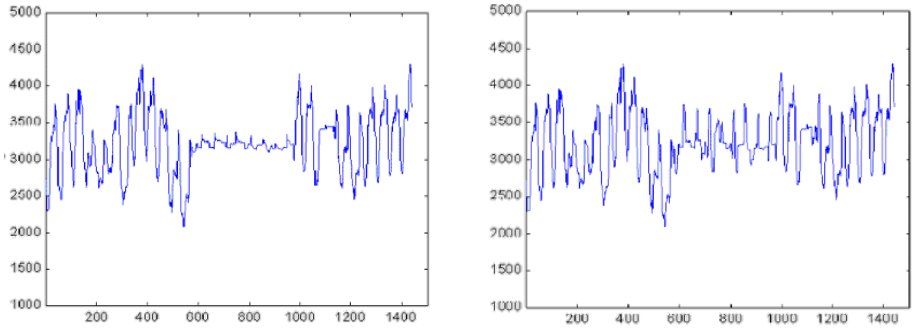


Figure 5.3 – Aggregated 30-to-30 minute Load Series. Month: September.
(a) Original Series (b) Filtered Series.

This example was meant only as a qualitative illustration that the methodology is well behaved. In the next section we shall quantitatively evaluate the performance of the procedure in a variety of situations.

6. Test and Results

To evaluate the performance of the presented methodology, we have included the results of simulations using a sample of three months (December 1999, January 2000 and March 2000) of minute-by-minute load series of two Brazilian utilities in the Southeast and South region. We selected from each one of these utilities samples consisting of two sets of days: 10 ordinary days (Monday to Sunday, without holidays or special events) and 6 days of special events and holidays. For these sets, artificial irregularities were introduced in the series in order to evaluate a posteriori the agreement of the values obtained by filtering. We conducted four types of test: (i) one for outlier detection, two different missing value interpolation tests using (ii) short and (iii) long segments of consecutives missing values, and (iv) a daily pattern test, as described below. These four tests were carried out for each utility (South, Southeast), for each of the two different types of day sets (ordinary days and special events or holidays) as well as for each of the two subgroups of load condition (peak and off-peak), totaling 16 tests per utility (32 tests). For test validation purposes, the MAPE statistics were considered (Mean Absolute Percentage Error) or the percentage of agreement, depending on the type of test. We consider MAPE more illustrative in our case since it allows us to compare between regions with different level of total load.

6.1 Outlier Test

For this test, outliers were artificially generated in the set of times described in Table 1. In observations chosen to become outliers, the value of the real load was increased by 50%, that is, if in a specific minute the value of the load were 1000, it would be substituted by 1500. The resulting data was then passed through the outlier filter. The table containing the data where outliers were generated can be found in Annex 2 of Table A1. The filter detected all outliers generated in this fashion, for a 100% success rate.

6.2 Missing Value Tests

Two types of missing values were considered: short and long term sets.

6.2.1 Test for short-term missing values

In the original data, three 15-minute-long sequences were declared missing (i.e, their load values were set to -999.99); they appear in Table 4. The same missing-value sequences were considered for both utilities. Table A2 in Annex 2 contains data from which missing values were generated.

The MAPE of the value filled-in by the filter versus the real value for each sequence of missing values was computed for each day, for special days, and for common days, at peak hours and at normal off-hours, as well as the General MAPE. The results appear in Table 6.1.

Table 6.1 – Results of filter – short-term missing values

	MAPE	
	South	**Southeast**
Peak-hour demand	0.0463	0.0846
Off-hour demand	0.0595	0.0739
Special Days	0.0604	0.0842
Common Days	0.0449	0.0792
GENERAL	0.0507	0.0811

The results obtained with the application of the filter in short-term sequences can be considered good in terms of MAPE, all values being below 0.1%. For common days as well as for special days, the maximum values of this measure were under 0.118% and 0.125%, respectively for South and Southeast utilities. Average values occurred between 0.04% and 0.061% for South utility and between 0.07% and 0.085% for Southeast utility.

6.2.2 Test for long-term missing values

Two missing value sequences were generated for the selected days. For the subgroup with normal off-hour load conditions, a 3-hour-long sequence from 1:15 p.m. to 4:14 p.m. was considered (both for the utilities companies and for the types of days). For the subgroup with peak-hour load conditions, a 2-hour-long sequence from 6:30 p.m. to 8:29 p.m. was considered (both for the utilities companies and for the types of days). All times declared missing were set to value -9999.99. The MAPE of the value filled by the filter in relation to

the real value was computed. These values can be found in Table 6.2. Besides the interval value, there appear values consolidated by day, by type of day and the general value as well.

Table 6.2 – Results of the filter – long-term missing values

	MAPE	
	South	**Southeast**
Peak-hour demand	0.0233	0.0476
Off-hour demand	0.0657	0.0562
Special Days	0.0429	0.0603
Common Days	0.0387	0.0455
General	0.0403	0.0510

The results obtained with the application of the filter continue to be good when long-term sequences are considered. Both for weekdays as well as for holidays, the maximum values for this measure were under 0.062% and 0.105%, respectively, for South and Southeast. Average values remained between 0.023% and 0.066% for South utility and between 0.045% and 0.061% for Southeast utility. Values for peak-hour demand were slightly better but the length of the sequence was smaller than in the off–hour demand period.

6.3 Test for Pattern Filtering

The data used for this test were previously filtered and added in a thirty-to-thirty minute series. Thus, these data do not present any other type of irregularity. To introduce an irregularity into the load standard, each day was substituted by its average, so that this day had an unexpectedly low standard deviation. The MAPE of the values filled by the filter in relation to the real values were computed. The procedures for attainment of the profiles of the daily curve of Southeast and South utilities are detailed in Annex 1.

The results obtained through the application of the PF in 30-by-30-minute sequences turned out well. Both for common days as well as for special days, the maximum values of this measure were under 0.3% and 0.26%, respectively, for South and Southeast utilities. Average values were placed between 0.14% and 0.13% for South and between 0.096% and 0.128% for Southeast utility.

7. Conclusions

In this article we describe a methodology for the filtering of time series with high frequency of observations, which was implemented to handle real data collection and transmission problems in the Brazilian electric power system. The methodology is illustrated using the August 1999 load series, which happened to present all the irregularities here addressed. The procedure was tested for various scenarios and for different types of irregularities using Brazilian Southeast and South utilities load series. The performance of the proposed approach from the experiments carried out produced an overall MAPE of less than 1%, though for special days and for the Southeast utility the MAPE was slightly above 1%, which is considered quite reasonable for the problem. These examples and test showed good results, validating our methodology for its intended application.

Table 6.3 – Results of Pattern Test Utilities

Day	South	Southeast
	MAPE x Day	**MAPE x Day**
24 – Dec – 1999	0.1202	0.1184
31 – Dec – 1999	0.0950	0.1136
01 – Jan – 2000	0.3008	0.0780
06 – Mar – 2000	0.0699	0.0847
07 – Mar – 2000	0.0896	0.0927
08 – Mar – 2000	0.1541	0.0909
07 – Dec – 1999	0.1283	0.1291
11 – Dec – 1999	0.0385	0.1298
12 – Dec – 1999	0.1923	0.0719
13 – Dec – 1999	0.1785	0.2561
16 – Dec – 1999	0.1051	0.1487
18 – Dec – 1999	0.0856	0.0684
19 – Dec – 1999	0.2204	0.0824
02 – Jan – 2000	0.2217	0.0532
04 – Jan – 2000	0.0918	0.2048
06 – Jan – 2000	0.0820	0.1346
MAPE Special Day	0.1383	0.0964
MAPE Common Day	0.1344	0.1279
Total MAPE	0.1359	0.1161

Acknowledgments

The authors would like to thank Sandra Canton for discussion of the ideas of this paper, Lucio de Medeiros and anonymous referees for their comments on the original draft of the paper, Marcelo Magnasco for the carefully revision of the manuscript and CNPq for financial support.

References

(1) Ameen, J.R.M. & Harrison, P.J. (1985). Normal Discount Bayesian Models (with discussion). **In:** *Bayesian Statistics 2* [edited by J.M. Bernardo *et al.*], North Holland, Amsterdam, 271-298.

(2) Brubacher, S.R. & Tunnicliffe Wilson, G. (1976). Interpolating Times Series with Application to the estimation of holiday Effects on electricity. *Applied Statistic*, **25**(2), 107-117.

(3) Canton, S. (1999). Detecção e substituição de descontinuidades nas séries de carga minuto a minuto do CNOS via fator de Bayes. Dissertação de Mestrado, DEE, PUC-Rio.

(4) Durbin, K. & Koopman, S.J. (2001). *Time Series Analysis by State Space Models.* Oxford, New York.

(5) Ferreiro, O. (1987). Methodologies for the estimation of missing observations in time series. *Statistics and Probability Letters*, **5**, 565-69.

(6) Gordon, F. (1996). Previsão de carga diária através de modelos estruturais usando splines. Doctoral Thesis, DEE/PUC-Rio.

(7) Harrison, P.J. (1965). Short term sales forecasting. *Applied Statistics*, **15**, 102-139.

(8) Harvey, A.C. & Pierse, R. (1984). Estimating missing observation in the economic time series. *Journal of American Statistical Association*, **79**(385), 125-131.

(9) Jeffreys, H.J. (1961). *Theory of Probability*. Third Edition. Clarendon Press, Oxford.

(10) Kohn, R. & Ansley, C.F. (1987). A New algorithm for spline smoothing based on smoothing a stochastic process. *SIAM J Sci Statistical Computing*, **8**, 33-48.

(11) Koopman, S.J.; Shephard, N. & Doornik, J.A. (1998). Statistical algorithms for models in state space using Ssf Pack 2.2. *Econometrics Journal*, **1**, 1-55.

(12) Koopman, S.J. (1991). Efficient smoothing algorithms for time series models. Doctoral Thesis.

(13) Koopman, S.J.; Shephard, N. & Doornik, J.A. (1998b). *SsfPack* 2.0: Statistical algorithms for models in state space. An *Ox* link to underlying C code. Available on website: cwis.kub.nl/~fews/center/staff/koopman/*SsfPack*.htm.

(14) Ljung, G.M. (1989). A note on the estimation of *missing values* in time series. *Commun. Statist. – Simula*., **18**(2), 459-465.

(15) Pole, A.; West, M. & Harrison, J. (1994). *Applied Bayesian Forecasting and Time Series Analysis*. Chapman & Hall, New York.

(16) Kass, R. & Raftery, A.E. (1995). Bayes factors. *Journals of the American Statistical Association*, **90**, 773-795.

(17) Rizzo, G.M. (2001). Previsão de Carga de Curtíssimo Prazo no Novo Cenário Elétrico Brasileiro. Master's Degree Dissertation, DEE, PUC-Rio.

(18) Sobral, A. (1999). Modelo de previsão horária de carga elétrica para Light. Master's Degree Dissertation, DEE, PUC-Rio.

(19) Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B*, **40**, 584-589.

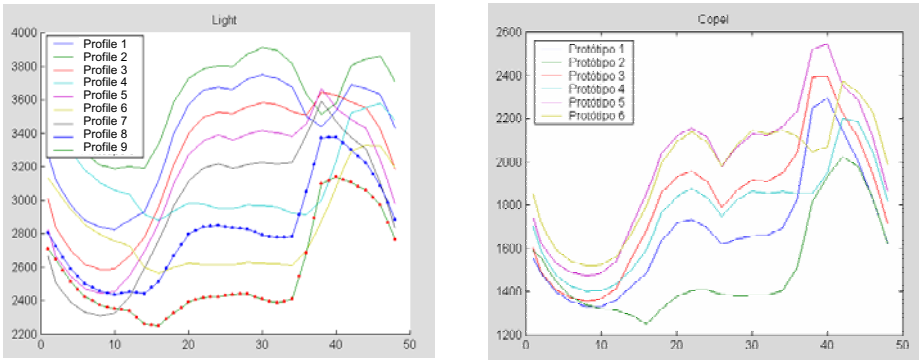
(20) Weinert, H.L.; Byrd, R.H. & Sidhu, G.S. (1980). A stochastic framework for recursive computation of spline functions: part II, Smoothing splines. *J. Optim. Theory Appl*., **30**, 255-268.

(21) West, M. & Harrison, J. (1989). *Bayesian Forecasting and Dinamic Models*. Springer-Verlag, New York.

(22) West, M. (1986). Bayesian Models Monitoring. *J. Roy. Statist. Soc. Ser. B*, **48**(1), 70-78.

Appendix 1 – Load Profiles for Southeast and South Utilities

The Load profiles used as base for the filter of standards was estimated by Sobral (1999) who used a neural network argument (Kohonen classification) to derive the different profiles for both utilities. The Kohonen algorithm produces 9 and 6 distinct profiles, respectively, for the two utilities being analyzed, as shown below in Figure 9. For more details, the reader is referred to Sobral (1999).



(a) (b)

Figure 9 – 30-to-30-minute-load-curve profiles.
a) Southeast utility b) South utility.

Appendix 2 – Observations Used in the Simulation

Table A1 – Observations that generated *outliers*

Type of Day	Year	Day	Day of the Week	Hours
Day of Special Event or holiday	1999	24 – Dec	Special Day	06:00; 15:00; 19:00
		31 – Dec	Special Day	06:10; 15:10; 19:10
	2000	01 – Jan	Holiday	06:20; 15:20; 19:20
		06 – Mar	Carnival	06:30; 15:30; 19:30
		07 – Mar	Carnival	06:40; 15:40; 19:40
		08 – Mar	Carnival	06:50; 15:50; 19:50
Normal Day	1999	07 – Dec	Tuesday	07:00; 16:00; 20:00
		11 – Dec	Saturday	07:10; 16:10; 20:10
		12 – Dec	Sunday	07:20; 16:20; 20:20
		13 – Dec	Monday	07:30; 16:30; 19:30
		16 – Dec	Thursday	07:40; 16:40; 20:40
		18 – Dec	Saturday	07:50; 16:50; 20:50
		19 – Dec	Sunday	08:00; 17:00; 21:00
	2000	02 – Jan	Sunday	08:10; 17:10; 21:10
		04 – Jan	Tuesday	08:20; 17:20; 21:20
		06 – Jan	Thursday	08:30; 17:30; 21:30

Table A2 – Short-term *missing values* sequence

Type of Day	Year	Day	Day of the Week	Hours
Day of Special Event or Holiday	1999	24 – Dec	Special Day	05:30 to 05:44; 14:30 to 14:44; 18:30 to 18:44
		31 – Dec	Special Day	05:35 to 05:59; 14:35 to 14:49; 18:35 to 18:59
	2000	01 – Jan	Holiday	05:40 to 05:54; 14:40 to 14:54; 18:40 to 18:54
		06 – Mar	Carnival	05:45 to 05:59; 14:45 to 14:59; 18:45 to 18:59
		07 – Mar	Carnival	05:50 to 06:04; 14:50 to 15:04; 18:50 to 19:04
		08 – Mar	Carnival	05:55 to 06:09; 14:55 to 15:09; 18:55 to 19:09
Normal Day	1999	07 – Dec	Tuesday	06:00 to 06:14; 15:00 to 15:14; 19:00 to 19:14
		11 – Dec	Saturday	06:05 to 06:19; 15:05 to 15:19; 19:05 to 19:19
		12 – Dec	Sunday	06:10 to 06:24; 15:10 to 15:24; 19:10 to 19:24
		13 – Dec	Monday	06:15 to 06:29; 15:15 to 15:29; 19:15 to 19:29
		16 – Dec	Thursday	06:20 to 06:34; 15:20 to 15:34; 19:20 to 19:34
		18 – Dec	Saturday	06:25 to 06:39; 15:25 to 15:39; 19:25 to 19:39
		19 – Dec	Sunday	06:30 to 06:44; 15:30 to 15:44; 19:30 to 19:44
	2000	02 – Jan	Sunday	06:35 to 06:49; 15:35 to 15:49; 19:35 to 19:49
		04 – Jan	Tuesday	06:40 to 06:54; 15:40 to 15:54; 19:40 to 19:54
		06 – Jan	Thursday	06:45 to 06:59; 15:45 to 15:59; 19:45 to 19:59