UMA METODOLOGIA PARA A FILTRAGEM DE SÉRIES TEMPORAIS. APLICAÇÃO EM SÉRIES DE CARGA ELÉTRICA MINUTO A MINUTO

Mayte S. Fariñas, Rodrigo Lage de Sousa e Reinaldo Castro Souza
Departamento de Engenharia Elétrica, PUC-Rio[1]

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 e o emprego do Fator de Bayes. Na interpolação de valores faltantes utiliza-se o Cúbico Suavizado. O desempenho da metodologia proposta é avaliado a través de vários testes onde as irregularidade foram simuladas.

Abstract

In this article a methodology for filtering time series is presented, with application to high frequency series. The goal of this methodology is to detect and to substitute the irregularities of the time series that can produce distortions on the modeling stage. The dynamic linear model and the Bayes factor tool, used to detect outlier values are detailed. To interpolate the missing values, the Smoothing Cubic is used. The performance of the methodology is evaluated using some tests where the irregularities had been artificially created.

Keywords:

[1] DEE PUC-RIO CP. 38063 Rio de Janeiro, CEP 22452-970
e-mails: {mayte, rsousa, reinaldo}@ele.puc-rio.br

Introdução

A exemplo do que ocorre mundialmente, o setor elétrico brasileiro vem passando por transformações profundas ao longo dos últimos anos. Diante deste novo contexto, a previsão de carga de curtíssimo prazo assume uma importância vital, servindo de base tanto para o cálculo do preço futuro da energia elétrica, quanto para a programação do despacho ótimo das usinas do sistema realizado pelo Operador Nacional do Sistema Elétrico (ONS) que se encarrega da operação do sistema.

Com o objetivo de desenvolver um protótipo de um sistema computacional de previsão das séries de carga elétrica a curto prazo para o ONS, foi realizado o Projeto Cahora, através de uma parceria entre o Cepel e a PUC-Rio. Este sistema e os modelos implementados encontram-se detalhado em Rizzo(2001). No entanto, o primeiro grande problema no desenvolvimento dos modelos de previsão foi a grande quantidade de irregularidades no sistema de aquisição de dados como descrevemos a seguir.

O CNOS, situado em Brasília, é o responsável pela consolidação das medições da carga das concessionárias de energia elétrica do sistema brasileiro. As leituras dos pontos de medição chegam ao sistema de aquisição de dados, via telemetria, a cada 20 (vinte) segundos aproximadamente. A cada minuto, é feito um registro do valor disponível na leitura. Este sistema de aquisição de dados está sujeito a ocorrência de irregularidades que podem ser motivadas pela transmissão de informação, erros na consolidação dos dados ou imprevistos no comportamento da carga ( ). Estas irregularidades podem ser descritas principalmente como:

Ocorrência de valores faltantes( ): quando não é realizada a leitura de nenhum valor de carga, durante o intervalo de um minuto.

Ocorrência de : quando o valor registrado é absurdo, discordante com o comportamento esperado. Isto pode ser devido por exemplo à uma falha na medição de algum ponto de intercâmbio ou mesmo a problemas inerentes ao processo de envio da informação. Incluem-se nesta categoria os valores reais, correspondentes a eventos inesperados e imprevisíveis como queda de carga, apagão.

Estas irregularidades prejudicam o desempenho de qualquer modelo de previsão que seja utilizado para prever as séries resultantes. Assim, antes de realizar a modelagem das séries de carga de meia hora, foi desenvolvido um módulo de tratamento de dados. A partir do Filtro de e (FMO) são corrigidas estas irregularidades transformando os registros de minuto a minuto em séries históricas modeláveis estatisticamente, que podem ser agregadas para trabalhar com freqüências maiores (por exemplo. meia em meia hora).

O filtro de têm duas funções principais, Detecção de , declarando-os como valores faltantes a Interpolação de todos os valores faltantes resultantes. Uma vez declarados os valores , esses valores devem ser substituídos por valores concordantes à série, para passar à etapa de modelagem da série de carga elétrica. Operacionalmente, ao final do dia, o sistema filtra os dados obtidos minuto a minuto.

Após a aplicação do FMO sobre as séries minuto a minuto algumas irregularidades ainda persistem nas séries filtradas. Estas irregularidades fujem aos padrões esperados. Assim, implementou um procedimento chamado de Filtro de Padrões(FP), para detectar e substituir estas irregularidades. O FP baseia-se na comparação das curvas diárias de carga com padrões estabelecidos via redes neurais.

Neste artigo se descreve a metodologia implementada para a detecção e interpolação de valores faltantes. São discutidos sucintamente aspetos teóricos do Modelo Linear Dinâmico e do fator de Bayes na detecção de irregularidades e do uso do Cúbico Suavizado na interpolação. Na seção 5 mostra-se através de um exemplo o esquema de funcionamento dos filtros. Na seção 6 são descritos os resultados obtidos nas simulações realizadas para avaliar o

desempenho da metodologia proposta.

2. Detecção de Discontinuidades.

Para detectar os valores da série de carga, utilizou-se um Modelo Linear Bayesiano com fator de desconto para a tendência e declividade e fator de escala para modelar a variância das observações como proposto em West (1986). Para a detecção das descontinuidades, utilizou-se o fator de Bayes, como descrito em West&Harrison(1986). Um modelo mais simples foi também utilizado por Canton(1999) para detecção de irregularidades. Como algumas descontinuidades importantes não se enquadravam a esta primeira implementação, como as irregularidades que serão descritas mais adiante, algumas modificações ao procedimento original foram feitas para detectá-las. A seguir descrevemos sucintamente o processo de detecção de .

2.1. O Modelo Linear Dinâmico (MLD)

Seja Y_t a série temporal, o MLD é descrito por:

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

onde t=1,...,T; Y_t é o vetor de observações da série no tempo t de dimensão nx1; θ_t é o vetor de parâmetros de dimensão kx1; contendo as componentes não observáveis do sistema (nível, declividade, sazonalidade, etc.); ν_t e w_t são erros aleatórios normalmente distribuídos, onde ν_t e w_t são não-correlacionados.

Na formulação Bayesiana deste modelo devem ser especificadas as variâncias dos erros ou seja, as matrizes V_t e W_t. O problema da evolução das variâncias dos erros deve ser solucionado, quando assumimos que elas são desconhecidas. Pole (1994), efetua o aprendizado da variância das observações V_t, em função do fator de escala ϕ que, no caso de normalidade, é a própria precisão da distribuição. A idéia é escalonar todas as variâncias e covariâncias das equações de atualização do MLD por ϕ. As equações recursivas do Filtro de Kalman são ainda válidas, porém as distribuições condicionais a ϕ serão normais enquanto que as marginais serão t de Student com os graus de liberdade da correspondente distribuição marginal de ϕ. Trabalhando-se com $V_t=\phi^{-1}$, o modelo (2.1) pode ser rescrito como:

$$\begin{aligned} Y_t &= F_t'.\theta_t + \nu_t \qquad \nu_t \sim N[0, \phi^{-1}] \\ \theta_t &= G_t.\theta_{t-1} + \omega_t \qquad \omega_t \sim N[0, W_t^*\phi^{-1}] \end{aligned} \tag{2. 2}$$

Para determinar a matriz W_t, será utilizada a formulação de fatores de desconto, introduzidos por Harrison (1965). A idéia baseia-se na consideração que o conteúdo informativo de uma observação decai com sua idade. Assim, a variância a priori no tempo t é calculada como função da variância a posteriori no tempo t-1, determinada por um fator de desconto δ ($0<\delta\leq 1$), que representa a quantidade de informação perdida com o avanço temporal da série. Isto é eqüivalente a estabelecer a variância dos erros do vetor de estado como $W_t=(\delta^{-1}-1)G_tG_{t-1}G_t$, onde C_{t-1} é a variância a posteriori do vetor de estado. Neste caso, consideram-se fatores de desconto diferentes para cada componente do vetor de estado, como em Ameen e Harrison (1985). Assim, a matriz de desconto seria uma matriz diagonal $\delta=\text{diag}(\delta_1,\delta_2,\ldots,\delta_n)$.

2.2 Filtro de Kalman

A seguir mostramos resumidamente, as equações recursivas do filtro de Kalman para este modelo. As distribuições condicionadas ao valor de ϕ são as mesmas que no MLD com variâncias conhecidas e as distribuições marginais são obtidas a partir delas.

Inicialização.

Para o instante inicial t=0, assume-se uma distribuição Normal para a posteriori do vetor de estado e a Gama para o fator de escala:

$$(\theta_0 \mid D_0, \phi) \sim N[m_0, C_0^* \phi^{-1}] \quad (\phi \mid D_0) \sim G[n_0/2, d_0/2]$$

Prioris no instante t

vetor de estado: a priori para θ_t condicional a ϕ segue uma normal: $(\theta_t \mid D_{t-1}, \phi) \sim N[a_t, R_t^* \phi^{-1}]$, onde $R_t^* \phi^{-1} = R_t$.

fator de escala: a priori para o fator de escala ϕ segue uma distribuição Gamma:

$$(\phi_t \mid D_{t-1}) \sim G[n_{t-1}/2, d_{t-1}/2], \text{ com média } S_{t-1}=n_{t-1}/d_{t-1}.$$

A priori incondicional (marginal) para o vetor de estado pode ser calculada a partir das duas distribuições anteriores obtendo uma distribuição t de Student com n_{t-1} graus de liberdade. $(\theta_t \mid D_{t-1}) \sim T_{n_{t-1}}[a_t, R_t]$, onde $R_t=R_t^*S_{t-1}$.

Assim, a_t e R_t são calculados como no modelo com variância conhecida, considerando W_t, calculada através dos fatores de desconto. Logo: $a_t = G_t m_{t-1}$, $R_t = \delta^* G_t C_{t-1} G_t' \delta^*$ onde $\delta^* = \text{diag}(\delta_1^{-1/2}, \delta_2^{-1/2}, ..., \delta_n^{-1/2})$ é a matriz de desconto.

Previsão 1 passo-à-frente

A distribuição preditiva condicional ao valor de ϕ é: $(Y_t \mid D_{t-1}, \phi) \sim N[f_t, Q_t^* \phi^{-1}]$ com $f_t = F_t' a_t$ $Q_t^* = 1 + F_t' R_t^* F_t$.

A distribuição incondicional (marginal) para a distribuição preditiva pode ser calculada a partir da distribuição anterior e da priori para ϕ obtendo-se uma distribuição t de Student com n_{t-1} graus de liberdade: $(Y_t \mid D_{t-1}) \sim T_{n_{t-1}}[f_t, Q_t]$, com $Q_t=S_{t-1}Q_{t-1}^*$. Este resultado é similar ao que é obtido com variância conhecida, se considerarmos como estimativa pontual para ϕ até o instante t-1o valor esperado da priori S_{t-1}:

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

Posteriori (Atualização)

fator de escala: A posteriori para o fator de escala é obtida via Teorema de Bayes como: $p(\phi \mid D_{t-1}, Y_t) \sim G[n_t/2, d_t/2]$, onde $n_t = n_{t-1} + 1$, $d_t = d_{t-1} + e_t^2/Q_t^*$ e e_t são as inovações no instante t.

vetor de estado: levando-se em consideração as distribuições condicionais $p(\theta_t \mid D_{t-1}, \phi)$ e $p(Y_t \mid D_{t-1}, \phi)$ que são normais, obtém-se a distribuição conjunta de Y_t e θ_t e a seguir, a distribuição condicional para θ_t dado todos os valores observados : $(\theta_t \mid D_t, \phi) \sim N[m_t, C_t^* \phi^{-1}]$ onde $m_t = a_t + R_t^* F_t e_t / Q_t^*$ e $C_t^* \phi^{-1} = \phi^{-1}\left(R_t^* - R_t^* F_t F_t' R_t^* / Q_t^*\right)$

A distribuição incondicional (marginal) é obtida por integração da densidade anterior $p(\theta_t \mid D_t) = \int p(\theta_t, \phi \mid D_t) d\phi$ obtendo-se assim novamente uma t de Student $(\theta_t \mid D_t) \sim T_{n_t}[m_t, C_t]$ onde $C_t = C_t^* S_t$ e S_t é a estimativa pontual para o fator de escala. Sob estas considerações, lembrando-se que $R_t=R_t^*S_{t-1}$ as equações podem ser rescritas como:

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

2.3 Fator de Bayes

O conceito de fator de Bayes foi desenvolvido por Jeffreys (1961), e constitui uma ferramenta muito útil para testes de hipóteses e seleção de modelo. O fator de Bayes não é mais do que a razão entre a verossimilhança do modelo padrão e a verossimilhança de um modelo alternativo, e no contexto das séries temporais é utilizado como medida da capacidade preditiva do modelo. Assim, uma ampla discussão, no contexto clássico e bayesiano com múltiplas referências e aplicações pode ser encontrada em Kass&Raftery(1995).

West (1986) propõe um esquema de monitoramento seqüencial do modelo, baseado no fator de Bayes. Este esquema, discutido também em Harrison&West (1986), valida continuamente a capacidade preditiva do modelo, sob a abordagem bayesiana. Como o objetivo é desenvolver uma ferramenta sensível as falhas locais do modelo, adapta-se o conceito de fator de Bayes. Seja A o modelo alternativo e $p_A(Y_t \mid D_{t-1})$ a sua distribuição preditiva, define-se então o fator de Bayes como:

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

Se considerarmos $\Omega_t(0)=1 \ \forall$ t, tem-se a seguinte expressão recursiva para $\Omega_t(k)$: $\Omega_t(k) = H_t.\Omega_{t-1}(k-1)$, onde H_t é o fator de Bayes baseado somente em y_t:

$$H_t = p(y_t \mid D_{t-1})/p_A(y_t \mid D_{t-1}).$$

Dada uma alternativa adequada, um valor para H_t menor que 1 e significativamente pequeno, é uma indicação da existência de descontinuidade na série. H_t indica a existência de um outlier potencial mais em casos de mudanças pequenas o graduais os fatores de Bayes individuais não são suficientemente pequenos para indicar falha no modelo. Faz-se necessário então a utilização conjunta da quantidade $\Omega_t(k)$ que avalia o modelo frente ao modelo alternatico nos k instantes pasados. Assim, valores pequenos de $\Omega_t(k)$ sugerem uma possível mudança no passado. O algoritmo, como proposto por West(1986) resume-se a seguir:

Passo 1. Calculamos seqüencialmente as quantidades L_t e ℓ_t definidas por: $L_t=H_t \min(1,L_{t-1})$ e o comprimento da descontinuidade: $\ell_t = \begin{cases} \ell_{t-1}+1; & \text{se} \quad L_{t-1} < 1 \\ 1; & \text{se} \quad L_{t-1} \ge 1 \end{cases}$

Passo 2. Determinar a descontinuidade

A. Se $H_t \ge \tau$, y_t é consistente com o modelo S, ir a B.

Senão ($H_t<\tau$), y_t é , ir a 3A.

B. Se $L_t \ge \tau$, sistema sob controle ir a 3B.

Senão, ($L_t<\tau$,) or ($\ell_t>Q$), mudança estrutural. ir a 3A.

Passo 3. Atualização

A. Rejeitar a observação y_t. Atualizar as equações do Filtro de Kalman incrementando a incerteza: $m_t=a_t$, $C_t= \delta C_t$

B. A seqüência do modelo é satisfatória. Atualização padrão das equações do Filtro de Kalman.

Quando $\ell_t=1$ a descontinuidade é do tipo "transiente". Um valor $\ell_t>1$ indica uma mudança estrutural que começou no instante t-ℓ_t+1. Quando $\ell_t>Q$, mesmo que $L_t>\tau$, podemos admitir que neste instante esteja ocorrendo uma mudança lenta, a partir do instante t-ℓ_t+1.

No modelo considerado, as distribuições preditivas seguem uma t de Student $(Y_t \mid D_{t-1}) \sim T_{n_{t-1}}[f_t, Q_t]$. Se considerarmos o modelo alternativo baseada no uso do "power

discounting" como em Smith(1979), chegaremos a considerar uma densidade preditiva mais difusa, isto é com a mesma média e variância e Qa_t tal que $Q_t=\rho^2 Qa_t$. É necessário ainda estabelecer os valores para τ (real positivo), ρ(real entre 0 e 1) e Q(inteiro). Uma discussão desta escolha pode ser encontrada em West&Harrison(1986) e West(1986). Seguindo suas sugestões trabalhamos com ρ=0.4 e τ=1.5, Q=6.

Pelas características próprias da série trabalhada, foram necessárias algumas alterações no modelo proposto por West(1986). Nele, quando estamos em presença de qualquer tipo de descontinuidade a adaptação do modelo deve ser feita de tal modo que permita os câmbios. Assim a matriz de variâncias da posteriori é multiplicada por um fator maior do que um. Este fator, pode ser uma matriz, permitindo uma inflação diferente para a variância a posteriori de cada componente. Isto também é aconselhado no caso de presença de .

Este algoritmo deve identificar e mudanças estruturais. Em nossos dados há existência de 2 e até 4 consecutivos, sem que isso signifique uma mudança estrutural. Se utilizarmos diretamente a metodologia proposta por West, o incremento da incerteza do modelo frente a um dificultará a detecção posterior de outros . Como o objetivo aqui não é a modelagem e sim a detecção de descontinuidades, o incremento da incerteza só será efetuado quando detectamos uma mudança estrutural, ficando inalterada quando um é detectado. Por outro lado, nossos dados apresentam intervalos grandes com , por esta razão, incrementos consecutivos na matriz de variâncias, fariam esta ir para infinito, inutilizando numericamente o sistema de detecção de . Assim, decidiu-se não incrementar a variância em presença de .

Pela natureza das observações, quedas estruturais bruscas não podem acontecer porque a carga não pode ter uma variação forte de um minuto para outro, Assim os valores outliers e as mudanças estruturais devem ser substituídos. Quando uma mudança estrutural é detectada e ela é permanente, são substituidos todos os valores numa vizinhança do inicio da queda.

3. Interpolação de *Missing values*

Várias metodologias para estimação de valores faltantes em séries temporais tem sido propostas na literatura. Têm sido amplamente utilizadas as alternativas paramétricas (Brubacher, S. R. (1976)), que utilizam um modelo estatístico para ajustar os dados, utilizando-o depois na interpolação. Ferreiro (1987), Harvey & Pierse (1984) and Ljung (1989) utilizam esta abordagem.

Uma primeira idéia na substituição dos valores faltantes e foi a utilização do modelo Bayesiano ajustado na etapa de declaração de outlier, substituindo estas observações pelas previsões um passo-à-frente feita no instante anterior. Contudo, sendo que existiam trechos de dados com muitas observações faltantes (ou declaradas outlier) consecutivamente, às vezes por mais de uma hora (60 observações) foi preciso abandonar estas técnicas pois nestes casos quando o horizonte de previsão cresce a previsão não oferece bons resultados.

A grande dificuldade nesta abordagem é que a identificação do modelo e a estimação dos parâmetros ficam comprometidas em presença de valores faltantes (Stoffer (1986)). Alguns autores propõem processos iterativos de estimação para robustecer o ajuste do modelo como em Pourahamdi (1988) e também o algoritmo EM, proposto por Little&Rubin (1987) tem sido aplicado na interpolação de valores ausentes. Nestes casos, os procedimentos iterativos tornam o procedimento computacionalmente custoso ficando inviável quando trabalhamos com alto volume de dados. Assim, decidimos utilizar o cúbico suavizado que já foi utilizado na interpolação de séries temporais por Gordon (1996) e no tratamento de valores faltantes em Koopman (1991) e Koopman et al (1998).

3.1 *Spline* Cúbico Suavizado.

O cúbico suavizado (SCS) é uma das técnicas não paramétricas que mais tem sido aplicadas nos problemas de interpolação em séries temporais (Ver Ferreiro (1987) e Koopman (1998)) com bons resultados. Este método tem sido utilizado em alguns pacotes comercias tais como o SsfPack (Koopman (1998b)).

A idéia subjacente no cúbico, é o ajuste de um polinômio de terceira ordem. Este ajuste se efetua através de um modelo de espaço de estado que representa o , utilizando o Filtro de Kalman Suavizado, como descrito em Kohn & Ansley (1987). Este tipo de modelo leva em consideração todos os valores disponíveis e não só três pontos.

Weinert (1980) definem de modo geral o conceito de suavizado, discutindo o problema e estabelecendo a existência e unicidade das soluções. Kohn e Ansley (1987) discutem o caso particular de polinomiais definido como a seguir:

Seja $t_1, t_2, \ldots t_N$, os valores de tempo onde o valor da função f é conhecida. O problema de interpolação por polinomial suavizado consiste na procura do polinômio f que minimize a função de perda:

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

onde L é o operador diferencial $L=d^m/dt^m$ e quando m=2, teremos o cúbico suavizado e a função de perda consistira na integral do quadrado da curvatura, sendo que a solução pode ser procurada de tal forma que $Y_{ti}=f(t_i)$. Wahba(1978), Weinert (1980) representam o como o limite da esperança condicional de um processo estocástico. Isto permite, uma vez obtida a representação de espaço de estado do modelo estocástico, estabelecer algoritmos eficientes de ajuste do cúbico suavizado, através das equações recursivas do filtro de Kalman suavizado(Kohn &Ansley (1987), Koopman(1991), Koopman (1998)).

No caso em que o número de valores faltantes consecutivos é pequeno, o cúbico suavizado oferece uma boa solução. Não entanto, quando existem muitos valores faltantes consecutivos a interpolar o não consegue refletir todas as nuanças da curva de carga, porque a solução e suave. Foi por isso que pensou-se em substituir o valor faltantes por uma combinação convexa entre o cúbico e um valor que refletisse o padrão da curva de carga.

Seja Y_t a série de carga minuto a minuto. Suponha que Y_t é um valor faltante no instante $t=t_o$. Dito valor faltante será substituído pela combinação linear e convexa dos valores S_t e TY_t. Isto é: $Y_{to}=\alpha TY_{to}+(1-\alpha)S_{to}$, onde S_t é o valor interpolado via cúbico suavizado e SY_t valor de carga correspondente ao mesmo instante de tempo do dia com padrão de carga mais parecido ao dia atual. Dada a curva de carga, o procedimento procura um trecho parecido ao intervalo [t-trecho,t-1] onde t é o instante correspondente a Y_t o valor a substituir. O critério utilizado para selecionar o trecho parecido é o do menor erro quadrático médio, depois de ajustar ambos intervalos para terem a mesma média.

4. Filtro de Padrões

Após a aplicação do Filtro de sobre as séries minuto a minuto algumas irregularidades ainda persistem nas séries filtradas. Estas irregularidades estão referidas aos padrões esperados. Acontece, por exemplo, que a curva de carga de certo dia não corresponde com o padrão esperado para esse dia, no entanto não existem que tenham influenciado, na agregação, para a existência desta anomalia. Para solucionar isto decidiu-se implementar um filtro de padrões sobre a série agregada de 30 em 30 minutos.

4.1 Metodologia do Filtro de Padrões.

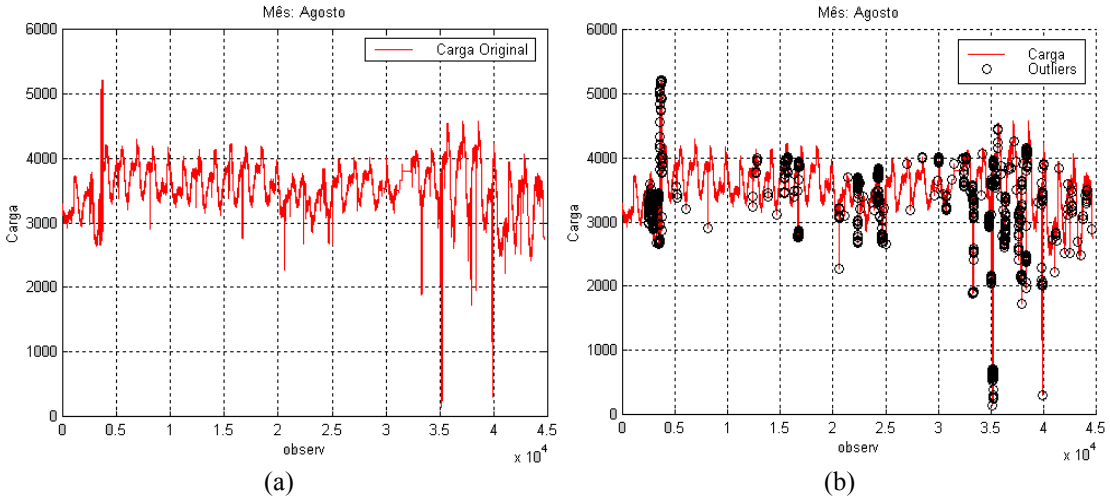
Em trabalho anterior, Sobral (1999) estabelece, através de uma rede neural de Kohonen (de 4x4), uma classificação dos tipos de curva de carga horária para a concessionária do Sudeste. Nesse trabalho, os dias foram agrupados segundo o padrão de carga horária e foram obtidos os seus respectivos protótipos. Estes protótipos servirão de base para a implementação do filtro de padrões.

Como a ideia é comparar os protótipos com a série de carga de 30 em 30 minutos, utilizaram-se os mesmos grupos, obtidos por Sobral(1999). Partindo dos protótipos da carga horária dos grupos, se obtém, via interpolação linear, os protótipos da curva de carga de meia em meia hora. Para cada dia d, temos a curva de carga de 30 em 30 minutos e devemos analisar se ela esta fora ou não do padrão esperado. Para isto seleciona-se, dentro do conjunto de protótipos, aquele que resulta mais adequado, considerando por exemplo, o que minimiza a estatística MAPE( ). Considerando que o objetivo aqui é avaliar o padrão de carga e não sua média, no cálculo do MAPE faz-se com que ambas curvas (a curva diária e o protótipo) tenham a mesma média. Se o MAPE do protótipo mais próximo é maior do que o limiar especificado, considera-se que esse dia está fora do padrão e deve ser substituído.

A substituição não se efetua no dia todo, porque pode acontecer que nem todos os intervalos do dia sejam responsáveis pela falha no padrão. Assim, divide-se o dia nos seguintes intervalos: 00:00h às 05:30h, 06:00h às 11:30h, 12:00h às 14:30h, 17:00h às 19:30h, 20:00h às 23:30h. Com eles, comprova-se quais tem o MAPE maior do que o limiar. Esses intervalos foram substituídos pelos respectivos intervalos da curva do protótipo ajustada para ter a mesma média que a curva do dia em estudo. O limiar escolhido para o MAPE foi de 5%. A primeira vista, este valor pode parecer pequeno demais, mais cabe lembrar que as duas curvas de carga (a do dia e a do protótipo) foram obrigadas a ter a mesma média. Para escolher esse valor, observou-se o comportamento desta estatística para vários dias incluindo alguns dias que tinham sido considerados pelos especialistas como fora de qualquer padrão.

5. Ilustrando o Funcionamento

Para mostrar o funcionamento da metodologia proposta para a detecção de e interpolação de valores faltantes, apresentamos os resultados obtidos, aplicados as séries minuto a minuto de uma concessionária do Sudeste. Pelo volume de dados apresentamos os dados correspondentes ao mês de Agosto de 1999, computando-se observações 44640 observações. Delas, 189 (0,42%) eram faltantes e 159 (3.04 %) foram declaradas .



(a) (b)

Figura 5.1- Série de Carga minuto a minuto da concessionária do Sudeste para o mês de Agosto (a) Série Original. (b) *Outliers* detectados.

Se utilizamos procedimento proposto para substituir os valores faltantes e da série de carga da concessionária do Sudeste para o mês Agosto, após a detecção de como na Figura 5.1, obteremos o seguinte resultado:

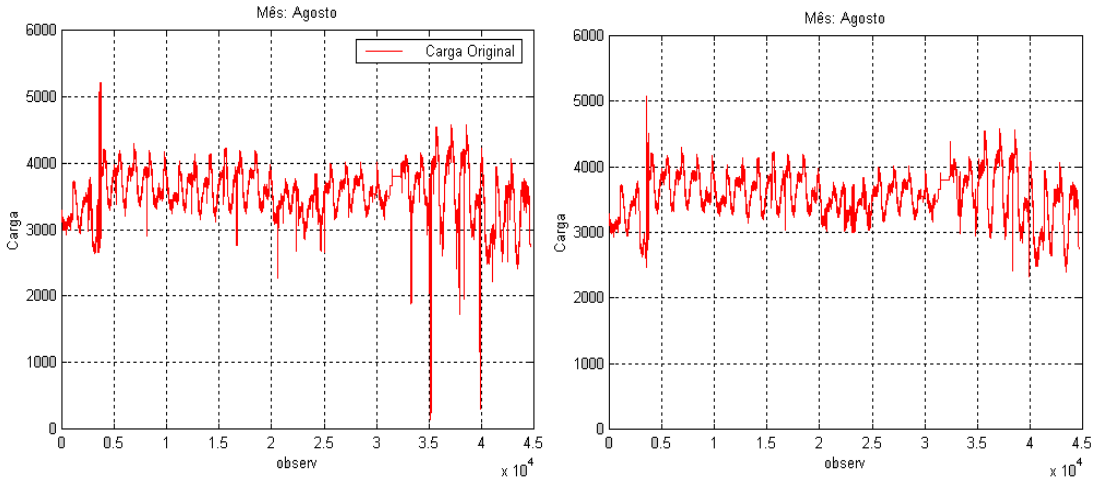
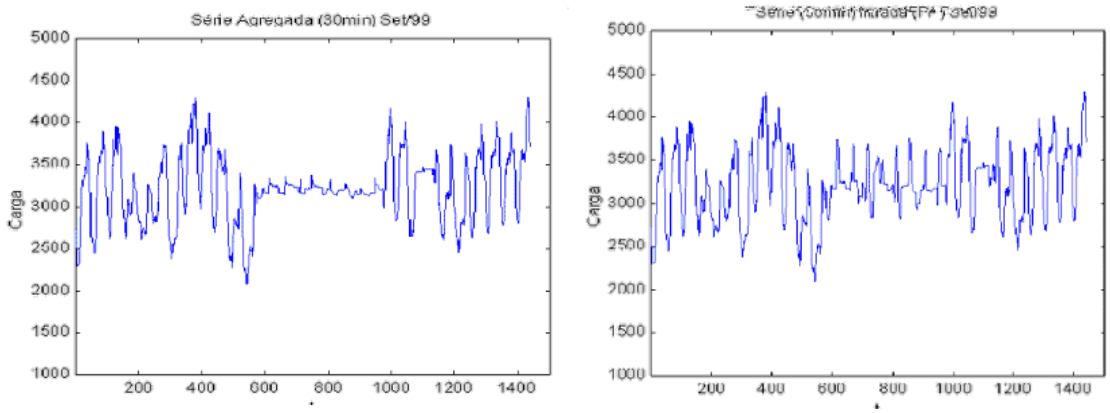


Figura 5.2- Série de Carga minuto a minuto para o mês de Agosto

(a) Série Original. (b) Série com os *outliers* e *missing* substituídos.

Enquanto a substituição dos valores pelo método proposto, é preciso apontar o fato que quando o intervalo de valores faltantes consecutivos começa a aumentar, o desempenho do cúbico na substituição dos valores faltantes piora. Isto se deve a que qunado o intevalo de valores é muito espaçado o spline cúbico ajustado apresenta uma curvatura mais acentuada, afastando-se do padrão esperado para a carga.



(a) **(b)**

Figura 5.3- Série de Carga Agregada de 30 minutos em 30 minutos para o mês de Setembro. (a Série Original, (b) Série Ajustada.

6. Testes e Resultados.

Para validarmos a metodologia para a filtragem de séries temporais foram realizados vários testes utilizando uma amostra base três meses de dados minuto a minuto de duas concessionárias, uma do Sudeste e outra do Sul. Foram selecionados da amostra de cada concessionária, 10 dias comuns (segunda a domingo sem feriados e sem eventos especiais) e 6 dias de eventos especiais e feriados, formando dois tipos de conjuntos de dias. Criaram-se artificalmente irregularidades em cada uma das série e foram realizados 1 teste de , 2 tipos de testes para valores faltantes e 1 teste de padrões conforme serão descritos. O experimento foi realizado separadamente para dois subconjuntos de condição de carga:

Condição de carga na ponta e fora da ponta.

Serão realizados três tipos de teste de valores faltantes e para as duas concessionárias, para cada um dos dois tipos de conjuntos de dias (dias comuns e dias de eventos especiais e feriados)e para cada um dos dois subconjuntos de condição de carga (na ponta e fora da ponta), num total de 24 testes. Na avaliação dos testes, considerou-se a estatística MAPE (Erro absoluto percentual médio) ou o percentual de acerto conforme o tipo de teste.

6.1 Teste de *Outliers*.

Para este teste, foram gerados artificialmente no conjunto de instantes de tempo descrito na tabela A1 (Anexo 2). Nas observações escolhidas para serem gerados, aumentou-se em 50% o valor da carga real, ou seja, se no minuto considerado o valor da carga era de 1000, este foi substituído por 1500, Os dados resultantes foram então passados pelo filtro de outlier. O filtro detectou os valores de em todos os instantes em que eles foram gerados. Logo o percentual de acertos foi igual a 100%.

6.2 Teste de *Missing Values*

Foram considerados dois tipos de valores faltantes: seqüências de curta duração e seqüências de longa duração.

6.2.1 Teste de *Missing* de curta duração

Nos dados originais, foram declaradas como missing três seqüências com 15 minutos de duração. As seqüências declaradas como valores ausentes, que assumiram valor de carga -999.99, aparecem na tabela A2 (anexo2). As mesmas seqüências de valores faltantes foram consideradas para ambas as concessionárias.

Foi computado o MAPE do valor preenchido pelo filtro em relação ao valor real para cada seqüência de valores faltantes, para cada dia, para dia especial, para dia comum, no horário de ponta e fora de ponta e o MAPE Geral. Os resultados aparecem na tabela 1.

Tabela 1: Resultados do filtro – missing de curta duração

	MAPE	
	Sul	Sudeste
Demanda na Ponta	0,0463	0,0846
Demanda Fora da Ponta	0,0595	0,0739
Dia Especial	0,0604	0,0842
Dia Comum	0,0449	0,0792
GERAL	0,0507	0,0811

Os resultados obtidos com a aplicação do filtro em seqüências de curta duração podem ser considerados bons, em termos de MAPE, todos menores que 0,1%. Tanto para dias comuns, quanto para dias especiais, os valores máximos dessa medida foram inferiores a 0,118% e 0,125%, para Sul e Sudeste, respectivamente. Os valores médios situaram-se entre 0,04% e 0,061% no Sul e entre 0,07% e 0,085% no Sudeste.

6.2.2 Teste de *Missing* de longa duração.

Foram geradas duas seqüências de valores faltantes para os dias selecionados. Para o subconjunto de condição de carga fora da ponta foi considerada uma seqüência com duração de 3 horas, sempre no intervalo de 13:15 a 16:14 (tanto para as concessionárias quanto para os tipos de dias). Para o subconjunto de condição de carga na ponta foi considerada uma seqüência

com duração de 2 horas, sempre no intervalo de 18:30 a 20:29 (tanto para as concessionárias quanto para os tipos de dias). Todas as seqüências declaradas como faltantes assumiram o valor –9999.99. Foram computados o MAPE do valor preenchido pelo filtro em relação ao valor real. Estes valores encontram-se na tabela 2 e além do valor por intervalo, aparecem consolidados por dia, por tipo de dia e o valor geral.

Tabela 2: Resultados do filtro – missing de longa duração

	MAPE	
	Sul	Sudeste
Demanda na Ponta	0,0233	0,0476
Demanda Fora da Ponta	0,0657	0,0562
Dia Especial	0,0429	0,0603
Dia Comum	0,0387	0,0455
GERAL	0,0403	0,0510

Os resultados obtidos com a aplicação do filtro continuam sendo bons quando consideraras seqüências de longa duração. Tanto para dias comuns, quanto para dias especiais, os valores máximos dessa medida foram inferiores a 0,062% e 0,105%, para Sul e Sudeste respectivamente. Os valores médios situaram-se entre 0,023% e 0,066% no Sul e entre 0,045% e 0,061% no Sudeste.

6.3 Teste de Padrões

Os dados utilizados para este teste foram previamente filtrados e agregados em série de trinta em trinta minutos. Assim, estes dados não apresentam nenhum outro tipo de irregularidade. Para introduzir uma irregularidade no padrão de carga, cada dia foi substituído pela sua média, a fim de que esse dia tivesse um padrão não esperado. Foram computados o MAPE dos valores preenchidos pelo filtro em relação aos valores reais. Os procedimentos para obtenção dos protótipos da curva diária das concessionárias estão detalhados no Anexo 1.

Tabela 3: Resultados das concessionárias

Dia	Sul	Sudeste
	Mape x Dia	Mape x Dia
MAPE Dia Especial	0,1383	0,0964
MAPE Dia Comum	0,1344	0,1279
MAPE Total	0,1359	0,1161

Os resultados obtidos com a aplicação do filtro de padrões em seqüências de 30 em 30 minutos mostraram-se bons. Tanto para dias comuns, quanto para dias especiais, os valores máximos dessa medida foram inferiores a 0,3008% e 0,2561%, para Sul e Sudeste, respectivamente. Os valores médios situaram-se entre 0,1383% e 0,1344% no Sul e entre 0,0964% e 0,1279% no Sudeste.

7. Conclusões

Neste artigo foi apresentada uma metodologia para a filtragem de séries temporais com alta freqüência de observações. A metodologia proposta foi implementada para solucionar um problema real do sistema elétrico brasileiro. Ilustra-se a metodologia utilizando séries de demanda elétrica de duas concessionárias, uma do Sudeste e outra do Sul. Os procedimentos foram testados via simulação para vários tipos de dias e de irregularidade. Os resultados avaliam a performance da metodologia proposta: em todos os casos os resultados obtidos apresentaram erros com MAPE menores do que 1%, sendo ligeiramente superiores para os dias especiais como era esperado e para a concessionária do Sudeste.

Referências:

Brubacher, S. R. and Tunnicliffe Wilson, G.(1976). Interpolating Times Series with Application to the estimation of holiday Effects on electricity . Vol. 25, No. 2, pp. 107-117

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

Ferreiro, O. (1987). Methodologies for the estimation of missing observations in time series , No. 5, pp. 565-69.

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

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

Harvey, A.C. and Pierse, R. (1984). Estimating missing observation in the economic time series. , 79(385):125-131.

Jeffreys, H.J. (1961). ; Third Edition. Ox-ford: Clarendon Press.

Kohn, R. e Ansley, C.F.(1987). A New algorithm for smoothing based on smoothing a stochastic process 8, 33-48.

Koopman, S. M., Shephard, N. and Doornik, J. A. (1998). Statistical algorithms for models in satate space using Ssf Pack 2.2 (1998), Vol.1, pp 1-55.

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

Koopman, S.J., Shephard, N. e Doornik, J. A. (1998b). 2.0: Statistical algorithms for models in state space. An link to underlying C code. Disponível na página web: cwis.kub.nl/~few s/center/staff/koopman/ .htm.

Litlle, R. e Rubin, D. (1987). . New York: Wiley.

Ljung, G.M. (1989). A note on the estimation of in time series ., No. 18(2), pp. 459-465

Pole, A., West, M. and Harrison J, (1994). . Chapman&Hall, New York.

Purahmadi, M. (1988). Estimation and interpolation of of a stationary time series Vol. 10, No. 2.

R. Kass, and A. E. Raftery(1995). Bayes factors, , vol. 90, pp. 773-795, 1995

Rizzo, G. M (2001). Previsão de Carga de Curtíssimo prazo no Novo Cenário Elétrico Brasileiro. Dissertação de Mestrado DEE, PUC-Rio.

Sobral A. (1999). Modelo de previsão horária de carga elétrica para Light. Disertação de Mestrado. DEE, PUC-Rio.

Stoffer, D.S. (1986). Estimation and identification of space-time ARMAX models in the presence of missing data. 81-395.

Wahba, G(1978). Improper priors, smoothing and the problem od guardiang against model errors in regression , 40, pp. 584-589.

Weinert, H. L., Byrd, R.H. e Sidhu, G.S. (1980). A stochastic framework for recursive computation of functions: part II , Smoothing , ., 30, pp. 255-268.

West, M e Harrison, J. (1986). Monitoring and adaptation in bayesian forecasting models Vol81, No395 Theory and Methods.

West, M e Harrison, J. (1989) . Springer-Verlag, New York.

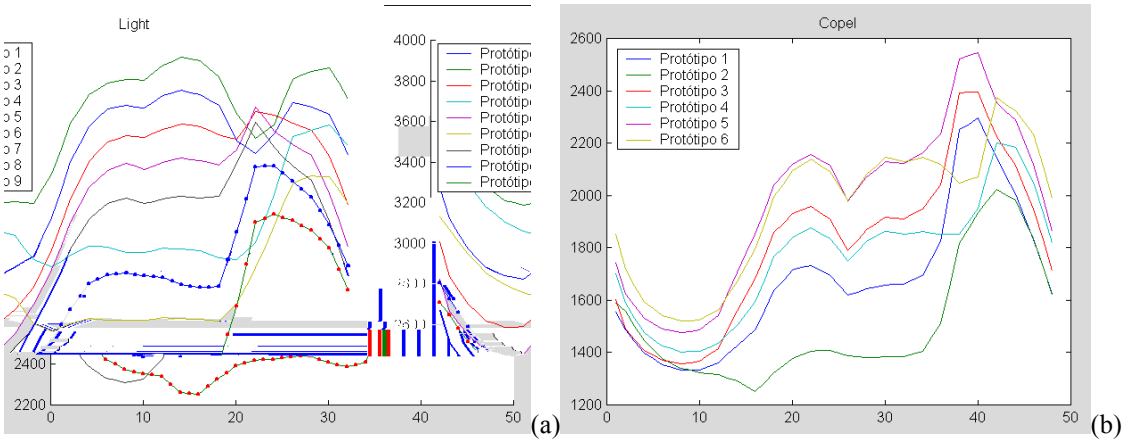
West, M. (1986). Bayesian Models Monitoring , Vol 48, No1, pp 70-78.

ANEXO 1. PROTÓTIPO DE CARGA PARA ÀS CONCESSIONÁRIAS.

Os protótipos de Carga utilizados como base para o filtro de padrões foram determinados segundo a metodologia proposta por Sobral(1999). No caso do Sudeste utilizaram-se os próprios resultados obtidos por Sobral(1999). Partindo de uma rede de Kohonen de 4x4, foram obtidos inicialmente 12 protótipos, reduzidos posteriormente para 9, segundo considerações discutidas na referência supracitada.

Seguindo esta metodologia, forma obtidos os protótipos para as curvas de carga horária do Sul, utilizando uma rede de Kohonen de 2x3, obtendo 6 protótipos. Os protótipos obtidos, para as curvas de carga horária, foram transformados em protótipos de curva de carga de 30 em

30 minutos a través de interpolação linear. Os protótipos são armazenados em arquivos, contendo matrizes de 9x48 e 6x48 para Sudeste e Sul, respetivamente, onde cada linha contem o padrão da curva de carga de 30 em 30 minutos para cada grupo o protótipo.



(a) (b)

Figura 9. Protótipos de curva de carga de 30 em 30 minutos. a) Sudeste, b)Sul.

ANEXO 2 OBSERVAÇÕES GERADAS PARA SIMULAÇÃO.

Tabela A1:Observações onde foram gerados os *outliers*

Dia de Evento especial ou feriado			Dia Comum		
Data	Dia da semana	Horas	Data	Dia da semana	Horas
24 – Dec/99	Dia Especial	06:00 ; 15:00; 19:00	07 – Dec/99	3a-Feira	07:00 ; 16:00; 20:00
31 – Dec/99	Dia Especial	06:10 ; 15:10; 19:10	11 – Dec/99	Sábado	07:10 ; 16:10; 20:10
01 – Jan/00	Feriado	06:20 ; 15:20; 19:20	12 – Dec/99	Domingo	07:20 ; 16:20; 20:20
06 – Mar/00	Carnaval	06:30 ; 15:30; 19:30	13 – Dec/99	2a-Feira	07:30 ; 16:30; 19:30
07 – Mar/00	Carnaval	06:40 ; 15:40; 19:40	16 – Dec/99	5a-Feira	07:40 ; 16:40; 20:40
08 – Mar/00	Carnaval	06:50 ; 15:50; 19:50	18 – Dec/99	Sábado	07:50 ; 16:50; 20:50