MONITORAÇÃO DA QUALIDADE DE DADOS EM SÉRIES TEMPORAIS DE CARGA ELÉTRICA UTILIZANDO REDES NEURAIS E ANÁLISE DE COMPONENTES INDEPENDENTES

José M. Faier e José M. de Seixas

Laboratório de Processamento de Sinais, COPPE/EP,Universidade Federal do Rio de Janeiro
Caixa Postal 68504, CEP 21845-970 Rio de Janeiro, RJ, BRASIL
faier@lps.ufrj.br ; seixas@lps.ufrj.br

Abstract – In this paper independent component analysis and neural processing are used for data quality monitoring in electrical energy time series. The independent component extraction increases the system performance, reducing the forecast error, adding relevant information and decreasing the validation corridor width. Therefore, the indpendent component analysis can be used as an additional tool for energy time series preprocessing, aiming at data quality monitoring.

Keywords – Data Quality, Time Series, ICA, Neural Networks, Data Quality Monitoring System.

Resumo – Este artigo propõe a aplicação de processamento neural sobre componentes independentes para monitorar a qualidade de dados em séries temporais de carga elétrica. Observou-se que a extração de componentes independentes melhorou o desempenho do sistema, estreitando os corredores de validação, agregando informações relevantes e fornecendo previsões de carga mais acuradas. Assim, a análise de componentes independentes pode ser utilizada como ferramenta adicional no pré-processamento de séries de energia, visando a qualidade de dados.

Palavras-chave – Qualidade de dados, Séries temporais, ICA, Redes Neurais, Sistema de Monitoração.

1. Introdução

Durante os últimos anos, o desenvolvimento mundial foi devido, em grande parte, à ampla disseminação de dados, principalmente pela difusão da Internet. De fato, com o crescimento no volume de dados, as atenções têm se voltado para a habilidade em absorver informações e responder adequadamente a elas [1]. Assim, o tema qualidade de dados tem se tornado cada vez mais evidente, pois é um fator fundamental para a transformação de dados em informações confiáveis.

Define-se qualidade de dados como o nível de correção, completude, consistência, interpretabilidade, segurança, informação agregada e outras características dos dados em conformidade com as especificações exigidas pelos usuários [2]. Assim, os bancos de dados devem ser monitorados continuamente em conformidade com as necessidades dos usuários em relação à qualidade [3].

Neste trabalho, é desenvolvido um módulo de monitoração da qualidade de dados para séries temporais com foco em séries de carga elétrica. Estas séries podem conter padrões fundamentais para a tomada de decisões e não podem estar corrompidas. Dessa forma, um sistema que monitore a qualidade, identificando problemas e, eventualmente, corrigindo falhas e agregando informações relevantes aos dados, conforme as especificações dos usuários, é de fundamental importância.

Para monitorar os aspectos fundamentais da qualidade neste tipo de série temporal, propõem-se corredores de validação para verificar cada amostra inserida no banco de dados e corrigi-la, caso necessário/solicitado. Os corredores são construídos dinamicamente com o auxílio de processos de predição neural. O centro do corredor é a previsão da amostra monitorada e as extremidades do corredor são obtidas a partir do erro de predição estimado. Este método permite a correção de dados corrompidos e faltantes, além de agregar informações relevantes para o processo de tomada de decisões.

Sistemas de processamento neural se mostram mais eficientes quando acompanhados por etapas de pré-processamento. Em trabalhos recentes com redes neurais, aplicadas ao processamento de séries temporais, alimenta-se a rede com o resíduo da série, após a etapa de pré-processamento [3]. O sistema de monitoração da qualidade proposto neste trabalho introduz a Análise de Componentes Independentes [4] para o auxílio do pré-processamento de séries de carga elétrica.

A Análise de Componentes Independentes (ICA) é uma técnica estatística e computacional para revelar fatores escondidos em conjuntos de sinais. A análise de componentes independentes define um modelo gerador dos dados multivariados observados, que são assumidos serem misturas de variáveis ocultas desconhecidas. As variáveis latentes são assumidas mutuamente independentes, e são chamadas de componentes independentes ou fontes dos dados observados. ICA tem sido utilizada como ferramenta auxiliar em processos auto-regressivos para a previsão de séries temporais [4]. O objetivo, com a inserção de ICA no pré-processamento dos dados, é obter fontes mais estruturadas e previsíveis, o que torna o sistema de monitoração da qualidade de dados mais eficiente.

Na próxima seção, é feita uma explanação da análise de componentes independentes aplicadas em séries temporais para a monitoração da qualidade de dados. Na Seção 3, apresenta-se o sistema de monitoração e, na Seção 4, o estudo de caso com séries de carga elétrica. O trabalho é finalizado na Seção 5, com as conclusões do estudo.

2. Análise de Componentes Independentes

A análise de componentes independentes (ICA) tem se mostrado importante no contexto de previsões de séries temporais [5]. ICA busca extrair as fontes a partir de misturas observáveis. Espera-se que estas fontes sejam tão independentes quanto possível e mais bem estruturadas do que os sinais observados.

Em sua modelagem básica, ICA busca sinais independentes utilizando estatística de ordem superior para estimar a matriz de separação B [4].

$$x=A*s \text{ e } y=B*x \tag{1}$$

onde,
x: sinais misturados observados;
s: fontes originais;
A: matriz de mistura;
y: fontes estimadas;
B: matriz de separação estimada.

Os algoritmos específicos para séries temporais se valem da informação existente na sua estrutura temporal. Neste caso, é possível utilizar informação de segunda ordem na obtenção dos componentes independentes [4].

Para se extrair as fontes em séries temporais, a informação da seqüência de amostras no tempo é utilizada. Assim, buscam-se fontes de modo que sejam descorrelacionadas entre si para diferentes atrasos no tempo. Neste sentido, uma das possibilidades é agrupar as estatísticas necessárias na Matriz de Auto-Covariância Cruzada atrasada:

$$C_\tau^x = E\left\{x(t)x(t-\tau)^T\right\} \tag{2}$$

onde,
x(t) são os sinais observados;
x(t-τ) são os sinais observados com atraso τ;

Aqui o ponto chave é que a informação contida em C_τ^x pode ser usada no lugar da informação de ordem superior. Assim, devemos encontrar uma matriz de separação (B) que faça, além da covariância instantânea (τ = 0) – branqueamento -, as covariâncias defasadas serem nulas, conforme equação abaixo:

$$E\left\{y_i(t)y_j(t-\tau)\right\}=0, i\neq j \tag{3}$$

onde, $y(t)=B * x(t)$ (4)

A motivação para se igualar a zero todas as covariâncias defasadas é o fato desta característica ser própria da independência [4].

Nessa metodologia, a informação de ordem superior é substituída por informação de segunda ordem, considerada em diferentes atrasos temporais. Assim, a matriz de auto-covariância atrasada fornece a informação necessária para se obter os componentes independentes. Alguns algoritmos de identificação cega de segunda ordem como, por exemplo, SOBI [6] e SOBI-RO [7], utilizam estes princípios e foram introduzidos no sistema de monitoração de qualidade de dados para séries temporais. Outros algoritmos, que utilizam os princípios gerais de ICA como, por exemplo, FastICA [8] e JADE [9], também foram incluídos no conjunto de ferramentas do sistema de monitoração.

3. Sistema de Monitoração da qualidade em Séries Temporais

O principal objetivo do sistema de monitoração da qualidade de dados em séries temporais é avaliar (e corrigir, caso necessário) a qualidade de uma nova amostra a ser incorporada ao banco de dados. Este sistema pode ser entendido sob o ponto de vista de um sistema de controle [3], no qual amostras passadas são utilizadas para compor o conhecimento da amostra atual e corrigi-la, caso necessário. No sistema desenvolvido, propôs-se a monitoração através de um corredor de validação baseado em métodos preditivos e suas respectivas incertezas. Assim, as amostras válidas da série temporal devem pertencer aos limites deste corredor. O objetivo é que o corredor seja o mais estreito possível, identificando os erros, corrigindo-os com boa acuidade, caso necessário/solicitado, e mantendo os valores considerados válidos.

A estrutura do sistema proposto é mostrada na Figura 1. A primeira fase do sistema de monitoração é a análise de componentes independentes. As fontes independentes $\boldsymbol{y}$ são obtidas a partir dos sinais observados $\boldsymbol{x}$. A incerteza no ordenamento das fontes [4], gerada pelo método ICA, pode afetar a construção dos modelos neurais. Para que o sistema seja imunizado deste problema, correlacionam-se as fontes independentes, obtidas e armazenadas durante a modelagem, com as fontes obtidas durante a operação do sistema. Dessa forma, as fontes podem ser ordenadas na operação tal como na modelagem. Em seguida, os componentes espúrios ou sem estrutura definida, como os ruídos, podem ser retirados (deflação). Assim, os sinais resultantes são compostos somente pelas fontes estimadas mais estruturadas.

A próxima etapa é a aplicação do pré-processamento (bloco pp - Figura 1) de forma a se retirar todo componente modelável [3]. Primeiramente, verifica-se a presença de heteroscedasticidade nas séries. Esta verificação pode ser direta, via inspeção visual, ou através um teste como o Goldfeld-Quandt [10]. Em caso de heteroscedasticidade, uma ação apropriada, como aplicação de função logarítmica, deve ser considerada. Com a série homoscedástica, é detectado se a tendência é estocástica ou deterministica. Para isto, utiliza-se uma combinação dos testes de Dickey-Fuller aumentado (ADF) [11] com o teste de Phillips-Perron [12]. O teste verifica se há raízes unitárias no processo de geração das séries temporais. Se houver, significa que a tendência é estocástica e a primeira diferença é aplicada n vezes (onde n é o número de raízes unitárias ou ordem de integração do processo) para torná-la estacionária. Se o teste não detecta raízes unitárias, pode-se concluir que a tendência é determinística. Neste caso, remove-se a tendência através de um ajuste polinomial.

Após a retirada da tendência, é verificada se há a presença de sazonalidades e de ciclos, através da análise espectral de Fourier e da análise da função de autocorrelação da série. Para se retirar os componentes mais significativos, realiza-se um teste de hipótese no espectro de Fourier para níveis de significância determinados pelo usuário. Detectados ciclos ou sazonalidades, estes são removidos pela retirada de componentes de freqüência do espectro.

Após esta etapa, o sistema verifica a correlação de todas as fontes (bloco CORR - Figura 1), incluindo seus respectivos atrasos, com a fonte alvo da previsão. Existindo correlações superiores a patamares previamente definidos pelo usuário, estas amostras são apresentadas para a rede neural (bloco RN - Figura 1) prever o resíduo do pré-processamento. Utilizam-se redes MLP [13] e Elman [14] para modelar o resíduo. A verificação da linearidade ou não linearidade da modelagem do resíduo é realizada utilizando critérios parcimoniosos [15], ou seja, verifica-se a hipótese de modelos mais simples capazes de modelar eficientemente as séries e, gradativamente, aumenta-se a complexidade do modelo.

O critério de parada do treinamento da rede baseia-se na capacidade de generalização do treinamento neural. Avaliam-se os erros médios quadráticos (EMQ) da saída da rede através de séries de treino e validação do modelo neural. O treinamento é paralisado quando são verificadas sucessivas elevações do erro com o conjunto de validação (número máximo de falhas) ou quando o número de épocas atinge o seu valor máximo definido pelo usuário.

A previsão neural do resíduo é remontada sobre a modelagem da série (bloco pp^{-1} – Figura 1) obtida pelo pré-processamento, resultando nas fontes estimadas ($\boldsymbol{y_{est}}$). Em seguida, o processo ICA é revertido (bloco ICA^{-1} – Figura 1) e a estimativa da amostra de interesse é finalmente obtida ($\boldsymbol{x_{est}}$).

No sistema proposto, o corredor de validação é a média do erro absoluto entre o valor estimado e o valor real da amostra (μ_{erro}), ajustado por uma constante k definida pelo usuário (Equação 5). O corredor é estimado utilizando-se o conjunto de treino/validação.

$$Corredor = k.\mu_{erro} \tag{5}$$

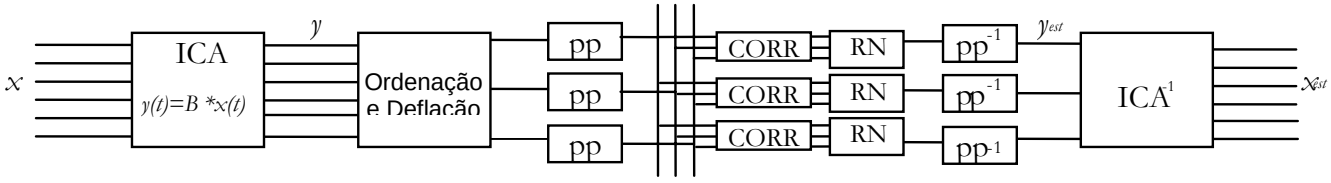

Figura 1. Estrutura do Sistema de Monitoração da Qualidade de Dados.

4. Estudo de Caso

Neste trabalho utilizam-se dados de carga e temperatura de uma concessionária de energia européia (*East-Slovakia Power Distribution Company*), a qual foi utilizada em uma competição promovida em 2001 pelo *European Network on Intelligent Technologies for Smart Adaptive Systems* [16]. Esta base de dados apresenta valores de carga em MW, coletados a cada trinta minutos, para o período de 1º de janeiro de 1997 a 31 de janeiro de 1999, e os valores médios diários de temperatura em °C, cobrindo o mesmo período da série de carga. Na competição realizada em 2001, a tarefa dos competidores foi desenvolver modelos para previsão do pico diário de carga para todo o mês de janeiro de 1999. Dessa forma, foram utilizados os valores dos anos de 1997 a 1998 para treino/validação e os dados de janeiro de 1999 para teste.

O conjunto de séries foi dividido em grupos compostos por séries relacionadas entre si. Para este estudo, utilizou-se o grupo com as 7 séries adjacentes ao horário de pico e a série de picos de carga diária. A série de temperaturas foi utilizada como série auxiliar. Monitoraram-se as séries relacionadas aos picos devido a sua importância no contexto de fornecimento de energia elétrica.

As séries são submetidas ao sistema de monitoração em sua forma bruta. Em seguida, são normalizadas e branqueadas [4] durante o processo ICA. A Análise de Componentes Independentes faz a extração cega das fontes, utilizando-se algoritmos de aplicação específica para séries temporais (SOBI e SOBI-RO) e os algoritmos de uso geral (FastICA e JADE) . Os melhores resultados com ICA (veja Tabelas 1 e 2) foram obtidos com o algoritmo de identificação cega de segunda ordem com ortogonalização robusta (SOBI-RO). Este algoritmo, antes de fazer a extração das fontes independentes, branqueia não apenas uma, mas uma combinação das matrizes de covariância atrasadas no tempo, para reduzir o efeito do ruído na modelagem ICA. Empiricamente, foram utilizadas 100 matrizes de covariância, referentes aos 100 primeiros atrasos da série.

As fontes independentes são pré-processadas segundo os testes de heteroscedasticidade, tendência, ciclos e sazonalidade (veja seção 3). Com exceção do teste no espectro de Fourier, no qual os melhores resultados utilizaram um nível de significância de 1%, todos os outros testes são realizados ao nível de significância de 5%.

Para a modelagem neural, utilizaram-se as redes MLP [13]. Dos valores entre 1997 e 1998, 70% foram utilizados para o treinamento e 30% para a validação do modelo neural. A topologia da rede é composta por 3 camadas: entrada, escondida e saída. Na camada de entrada, as correlações superiores a 5% definem as séries explicativas e os respectivos atrasos. Para a camada escondida, os testes de hipótese

ao nível de significância de 5% indicaram entre 1 e 2 neurônios. A função de ativação utilizada na camada escondida foi a tangente hiperbólica. Na camada de saída, apenas um neurônio com função de ativação linear foi utilizado.

As metodologias foram comparadas com base em dois indicadores: o erro absoluto de previsão do conjunto de teste ajustado pela constante k previamente definida durante o treino/validação, para que todas as amostras pertencessem ao domínio do corredor (Corredor Médio Absoluto - CMA), e o percentual de erro médio absoluto entre valores estimados e reais – MAPE (veja Equação 6 e 7).

$$CMA = k.\frac{\sum_{i=1}^{T}\left|\hat{x}_i - x_i\right|}{T} \tag{6}$$

$$MAPE = \frac{\sum_{i=1}^{T}\left|\frac{\hat{x}_i - x_i}{x_i}\right|}{T} \text{ x } 100 \tag{7}$$

onde, x_i e $\hat{x}_i$ são o sinal observado e sua predição no instante i, respectivamente, e T é o tamanho da janela temporal.

Na Tabela 1, são mostrados os valores do MAPE e do CMA, para os casos com e sem aplicação de ICA como parte da fase de pré-processamento. Com exceção do MAPE das séries 2 e 6 e o corredor da série 6, de uma maneira geral, utilizando-se ICA, houve uma redução significativa tanto do erro de previsão quanto do corredor de validação.

Avaliando-se isoladamente a série de picos, por ser uma série importante no contexto de monitoração de séries da carga elétrica, observou-se que o processamento ICA também resultou em um ganho no processamento para qualidade de dados. O CMA foi reduzido de 126 MW para 54 MW e o MAPE reduziu de 7,03% para 2,46% (veja Tabela 2). Os resultados mostraram que o modelo de monitoração com ICA, além de corrigir os dados corrompidos e agregar informação relevante para o usuário, permite a construção de corredores de validação mais acurados. Na comparação com os modelos da competição, o sistema proposto é capaz de monitorar a qualidade tão bem quanto os modelos mais bem colocados da competição. Respectivamente, o MAPE dos três primeiros colocados foi 1,98%, 2,14% e 2,49%.

Tabela 1 –CMA e MAPE de modelos com ICA (SOBI-RO) e sem ICA , para o grupo de séries entre 18h30min e 21h30min.

Série	Indicador	C/ ICA	S/ ICA
Série 1 (18h30min)	CMA (MW)	**129,5**	173,3
	MAPE (%)	**5,4**	9,5
Série 2 (19h)	CMA (MW)	**111,9**	136,4
	MAPE (%)	4,8	**4,6**
Série 3 (19h30min)	CMA (MW)	103,2	167,7
	MAPE (%)	**4,9**	9,0
Série 4 (20h)	CMA (MW)	**112,9**	146,3
	MAPE (%)	**5,3**	7,8
Série 5 (20h30min)	CMA (MW)	**105,6**	132,3
	MAPE (%)	**4,5**	7,6
Série 6 (21h)	CMA (MW)	109,0	**94,8**
	MAPE (%)	4,2	**3,4**
Série 7 (21h30min)	CMA (MW)	**94,7**	111,2
	MAPE (%)	**4,9**	5,7

Tabela 2 – CMA e MAPE de modelos com ICA (SOBI-RO) e sem ICA , para a série diária de picos de carga

Série	Indicador	C/ ICA	S/ ICA
Série diária de Picos	CMA	**54,0**	126,0
	MAPE (%)	**2,46**	7,03

Com a aplicação de ICA, a estrutura das séries de carga foi concentrada em poucas fontes independentes, permitindo uma modelagem neural mais acurada. Por exemplo, na Figura 2, são mostradas as séries de temperatura, a série de pico de carga e as fontes estimadas a partir destas duas séries temporais. Observa-se a diferença visual entre as duas fontes estimadas, indicando que as características presentes nas séries de temperatura e pico foram mapeadas para uma fonte apenas. Este mesmo comportamento foi observado no grupo das outras séries avaliadas. Estas fontes mais estruturadas contribuem para que o sistema neural seja mais eficiente na monitoração da qualidade dos dados. Ainda, a visualização das séries pode agregar informação aos usuários para a modelagem das séries. Por exemplo, algumas fontes, como os ruídos, poderiam ser retiradas após uma inspeção visual.
Na Figura 3 são mostrados os corredores de validação, a série de picos de carga e sua predição. Os corredores de validação auxiliam na detecção dos problemas de dados faltantes, corrompidos, desatualizados e *outliers*.

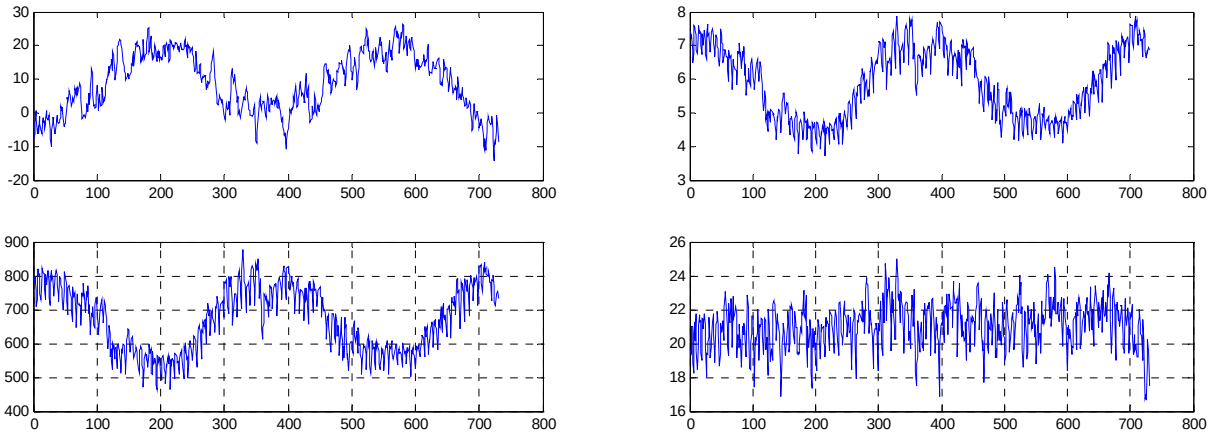

Figura 2 – À esquerda, acima a série de temperaturas e abaixo a série de picos de energia. À direita, as fontes independentes obtidas com ICA.

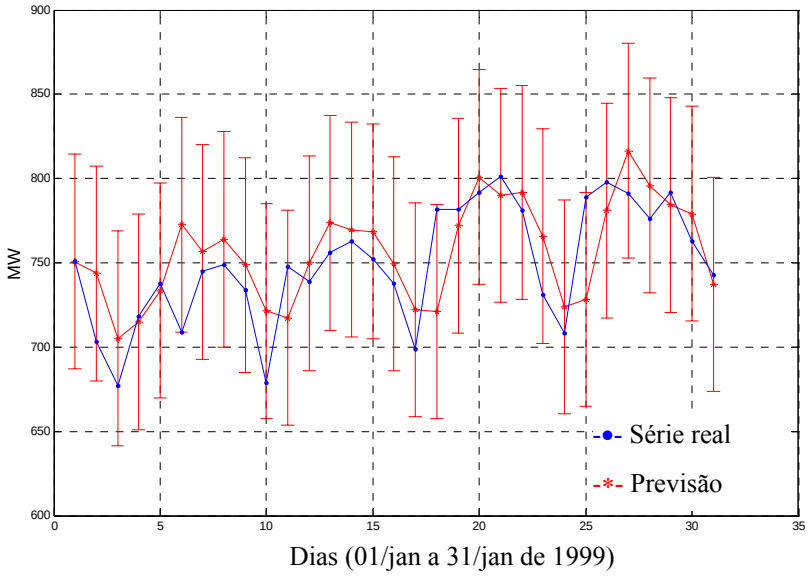

Figura 3 –Série de picos de carga, previsões e corredores de validação.

5. Conclusões

Para séries temporais propôs-se a monitoração da qualidade de dados baseada na construção de um modelo que permite avaliar novas amostras a serem incorporadas no banco de dados. Neste sistema, o modelo é composto por corredores de validação construídos a partir do erro de previsão.

Os corredores dinâmicos adaptam-se continuamente às variações estatísticas da série e o sistema emite um alerta para o usuário, quando o corredor não qualifica uma nova amostra.

No sistema proposto, redes neurais foram aplicadas de forma híbrida, em conjunto com uma fase de pré-processamento que inclui a análise de componentes independentes, no intuito de verificar o impacto da inserção de ICA no sistema de monitoração da qualidade de dados, particularmente para séries de carga elétrica. O modelo neural operou sobre fontes independentes, extraídas com ICA, e pré-processadas (com retirada de heteroscedasticidades, tendências, ciclos e sazonalidades). De uma maneira geral, o sistema de monitoração proposto, com o auxílio de ICA, reduziu a largura dos corredores de validação - o que tem impacto positivo na detecção de *outliers* – e, assim, pode auxiliar de forma acurada na substituição dos dados eventualmente corrompidos.

Agradecimentos: Agradecemos ao CNPq e à FAPERJ pelo suporte a este projeto e a Augusto Dantas (UFRJ) pelas discussões frutíferas realizadas no decorrer do desenvolvimento deste trabalho.

Referências:

[1] Eckerson, W. W. (2002). Data Quality and the Botton Line, Report, The Data Warehousing Institute.

[2] Chrisman, N. R. (1983). The Role of Quality Information in the Long-Term Functioning of a GIS. In: Proceedings of the AUTOCART06, v. 2, pp. 303-321.

[3] Dantas, A. C. H., Seixas, J. M. D. (2007). Neural Networks for Data Quality Monitoring of Time Series. In: 9th International Conference on Enterprise Information Systems.

[4] Hyvarinen, A., Karhunen, J. e Oja, E., (2001). Independent Component Analysis, ISBN 0-471-40540-X John Wiley & Sons, Inc.

[5] Kiviluoto, K., Oja, E. (1998). Independent Component Analysis for Parallel Financial Time Series. In Proc. Int. Conf. on Neural Information Processing (ICONIP'98), v. 2, pp. 895-898, Tokyo, Japan.

[6] Belouchrani, A., Abedi-Meraim, K., Cardoso, J., Moulines, E. (1997). A Blind Source Separation Technique Using Second Order Statistics. IEEE Transactions on Signal Processing, 45 (2):434-444.

[7] Belouchrani, A., Cichocki, A. (2001). Robust whitening procedure in blind source separation context, Electronics Letters, Vol. 36, No. 24, pp. 2050-2053.

[8] Hyvärinen, F. A. (1997). "A family of fixed-point algorithms for independent component analysis". In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97), p. 3917–3920, Munich, Germany.

[9] Cardoso, J. F., Souloumiac, A. (1993). Blind beamforming for non Gaussian signals.IEE Proceedings-F, 140(6):362–370.

[10] Rodrigues, S. A., Diniz, C. A. R. (2006). Modelo de Regressão Heteroscedástico, Revista de Matemática e Estatística, v. 24, n. 2, pp. 133-146.

[11] Dickey, D. A. and Fuller, W. A. (1979). Distributions of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, v. 75, pp. 427-431.

[12] Phillips, P. C. B. (1987). Time series regression with a unit root. Econometrica, v. 55, n. 2, pp. 277-301.

[13] Haykin, Simon. (1999) Neural Network: A Comprehensive Foundation, 2da. Edition ISBN 0-02-352761-7 Prentice Hall.

[14] Elman, J. L. (1990). Finding structure in time. Cognitive Science, v. 14, pp. 179-211.

[15] Medeiros, M. C., Teraasvirta, T., Rech, G. (2006). Building Neural Network Time Series Models: A Statistical Approach, Journal of Forecasting, v. 25, n. 1, pp. 49-75.

[16] EUNITE, European Network on Intelligent Tech-nologies for Smart Adaptive Systems (2001), http://neuron.tuke.sk/competition.