X Congresso Brasileiro de Inteligência Computacional (CBIC'2011), 8 a 11 de Novembro de 2011, Fortaleza, Ceará

Processo Estocástico Neural Aplicado em Séries de Afluências Mensais

Luciana Conceição Dias Campos
Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora
Juiz de Fora, MG, Brasil
luciana.campos@ufjf.edu.br

Marley M.B.R. Vellasco and Juan G.L. Lazo
Departamento de Engenharia Elétrica
Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro, RJ, Brasil
{marley,juan}@ele.puc-rio.br

Resumo – O modelo genérico de processo estocástico baseado em redes neurais, denominado Processo Estocástico Neural (PEN), foi aplicado no tratamento de séries de afluências mensais. Estas séries correspondem à Energia Natural Afluente (ENA), que é a estimativa da energia que pode ser gerada com todas as vazões afluentes a cada um dos reservatórios que compõem um reservatório equivalente de um subsistema do Sistema Interligado Nacional (SIN). Essas séries de ENA apresentam correlação temporal e correlação espacial. O modelo PEN na sua versão original pode capturar a correlação temporal, no entanto, não incorpora a correlação espacial dessas séries. Este trabalho apresenta uma variante do modelo PEN visando a incorporação da correlação espacial das séries de ENA. Os resultados indicaram que o modelo é capaz de capturar o comportamento da série temporal de todos os subsistemas do SIN, proporcionando diferentes cenários para os próximos 5 anos, que acompanham a mesma correlação temporal e espacial dos dados históricos.

Palavras-chave – Redes neurais, processo estocástico, séries de afluência mensais.

Abstract – The generic model of stochastic process based on neural networks, called Neural Stochastic Process (NSP), was applied to the treatment of series of monthly inflows. These series correspond to Affluent Natural Energy (ANE), which is the aggregation of the inflows to the plants, comprising a reservoir equivalent of a subsystem of National Interconnected System (NIS). The series of ANE presents temporal correlation and spatial correlation. The NSP model in its original version can capture the temporal correlation, however, does not incorporate the spatial correlation of the series. This paper presents a variant of the NSP model aimed at the incorporation of spatial correlation of the series of ANE. The results indicated that the model is able to capture the behavior of the time series of all NIS subsystems, providing different scenarios for the next 5 years that embody the same temporal and spatial correlation of the historical data.

Keywords – neural network, stochastic process, monthly inflows.

1. Introdução

O Sistema Interligado Nacional (SIN) é um sistema de coordenação e controle, formado pelas empresas das regiões Sul, Sudeste, Centro-Oeste, Nordeste e parte da região Norte, o qual congrega o sistema de produção e transmissão de energia elétrica do Brasil, que é um sistema hidrotérmico de grande porte, com predominância de usinas hidrelétricas [1].

Atualmente, o SIN é segmentado em quatro subsistemas correspondentes aos sistemas interligados: Sul, Sudeste/Centro-Oeste, Nordeste e Norte. Por se tratar de um sistema de grande porte, nos planejamentos de médio e de longo prazo, ocorre uma agregação dos reservatórios das usinas em reservatórios de energia equivalente, um para cada subsistema. Ocorre também a agregação das afluências às usinas em Energias Naturais Afluentes (ENA), que correspondem à estimativa da energia que pode ser gerada com todas as vazões afluentes a cada um dos reservatórios que compõem aquele reservatório equivalente, segundo uma política de operação [2].

As ENAs são séries não estacionárias, devido aos períodos de cheia e seca do ano, e sazonais com períodos de 12 meses, as quais, geralmente, apresentam correlações periódicas [3]. No contexto do planejamento da operação energética do sistema hidrotérmico brasileiro, a geração e análise de diferentes cenários é crucial para a obtenção de taxas de risco aceitável no futuro [2]. Pois o único cenário disponível na prática, que é o registro de afluências observadas (chamada série histórica), é insuficiente para compor uma amostra de tamanho necessário para estimar índices de risco aceitáveis [4]. Através da criação de cenários plausíveis, é possível reproduzir as características básicas dos dados históricos, extraindo melhor as informações dessa série temporal, permitindo a avaliação de riscos e incertezas pertinentes a um sistema hidroelétrico.

Para modelar o comportamento intrinsecamente não-lineares e estocásticos das séries de afluências mensais, um novo modelo genérico de processo estocástico [5] foi proposto em [6], chamado Processo Estocástico Neural (PEN). O PEN capta o comportamento da série histórica para gerar séries sintéticas, igualmente provável a série histórica, aplicável para a solução de diferentes tipos de problemas como os que envolvem fenômenos climáticos ou econômicos.

As ENAs apresentam correlação temporal (indicando a dependência com um evento que ocorreu anteriormente) e correlação espacial (que indica o quanto um evento em um local depende do que está acontecendo em outros lugares). O modelo PAR(p) [1], atualmente empregado pelo setor elétrico no tratamento das séries de ENA, fornece um modelo auto-regressivo para cada período da série [8]. Uma importante característica deste modelo é que ele preserva não apenas a correlação temporal das afluências mensais, mas também a correlação espacial entre os subsistemas do SIN [2, 7].

O modelo PEN na sua formação original é capaz de capturar a correlação temporal, no entanto, foi desenvolvido um modelo PEN diferente para a série de ENA de cada subsistema. Portanto, a formulação original não contempla a correlação espacial da série, ou seja, o ruído utilizado na geração de séries sintéticas de ENA são individuais por mês e por subsistema. Este trabalho apresenta uma variação do modelo PEN com o objetivo de incorporar a correlação espacial, além da correlação temporal, nos ruídos utilizados na geração de séries sintéticas de ENA.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta a descrição da nova versão do modelo PEN para o tratamento da série da ENA. A Seção 3 descreve a implementação deste novo modelo para gerar séries sintéticas de afluências mensais e a Seção 4 apresenta os resultados obtidos neste estudo de caso específico. Finalmente, a Seção 5 apresenta as conclusões do documento.

2 Descrição do Modelo PEN

A fim de manter a correlação espacial entre os subsistemas do SIN, decidiu-se implementar uma variação do modelo original PEN [6]. A idéia é deixar as redes neurais (RNA) do processo estocástico neural aprender a correspondência entre os subsistemas. Por esta razão, o novo modelo neural é um processo estocástico único para todos os subsistemas do SIN e suas RNAs têm quatro saídas, onde cada saída fornece valores para compor a série sintética de cada subsistema. Então, as séries de quatro subsistemas do SIN são geradas simultaneamente, de modo que este novo modelo é chamado de PEN$4out$.

O modelo PEN$4out$ mantém a estrutura baseada em redes neurais *Multilayer Perceptron* (MLP) [9] com uma única camada oculta, que são treinadas usando o algoritmo de aprendizado supervisionado *Levenberg-Marquardt* [10], uma variação do algoritmo *backpropagation* [9, 11].

Seja $Z(t)$ a série temporal de um subsistema. O índice de tempo t é então descrito pela equação 1:

$$t = (r-1) \cdot s + m \tag{1}$$

onde

- $r = 1 \ldots n$ é o ano e n é o número total de anos considerados pela série;
- $m = 1 \ldots s$ corresponde ao mês;
- $s = 12$ é o total de períodos da série;
- $n \cdot s$ é o tamanho da série observada.

Para modelar a série de ENA dos quatro subsistemas do SIN, o modelo PEN$4out$ é composto por 12 componentes estocásticas (CE), uma CE específica para cada mês m da série. Assim, a rede neural que compõe uma das CEs do modelo apresenta quatro saídas, sendo que cada saída representa um subsistema do SIN e, portanto, fornece valores para a série sintética deste subsistema que ela representa.

A Figura 1 ilustra a RNA do mês de janeiro, pois as configurações de todos os outros meses são semelhantes. Note que a RNA recebe como entradas quatro tipos de janelas temporais, uma para cada subsistema do SIN. Cada janela temporal é formada por valores passados da série de ENA do subsistema ao qual representa. O número de termos passado pode ser definido de forma independente para cada subsistema, no entanto, o estudo de caso apresentado foi realizado com janelas idênticas. Isto é, uma vez que definiu-se que a RNA do mês m tem ordem p_m , todas as quatro janelas temporais usam p_m termos passado da série que elas representam.

Seja Z_1 a série do subsistema Sudeste/Centro-Oeste, Z_2 a série do subsistema Sul, Z_3 a série do subsistema Nordeste e Z_4 a série do subsistema Norte. Assim, para obter um valor para a série do subsistema $\gamma = 1, \ldots, 4$ no tempo t, $Z_\gamma(t)$, o modelo PEN$4out$ acessa a CE correspondente ao mês m, e a sua rede neural recebe os termos: $Z_\gamma(t-1), Z_\gamma(t-2), \ldots, Z_\gamma(t-p_m)$ para todos subsistemas γ. Além disso, para reforçar a aprendizagem de comportamentos periódicos de cada série, o valor da série no período anterior, correspondendo a $Z_\gamma(t-s)$, também é adicionado na entrada da RNA para cada subsistema γ.

A Figura 2 mostra, em detalhes, um neurônio pertencente à camada oculta da rede neural do mês m com ordem p_m. A saída deste neurônio é dada pela equação 2.

[1] O modelo Auto-Regressivo Periódico (PAR) é uma generalização do modelo auto-regressivo (AR) e é amplamente utilizado na literatura de séries temporais para captar a correlação periódica, e sua principal característica é a variação dos parâmetros de acordo com o período [7]

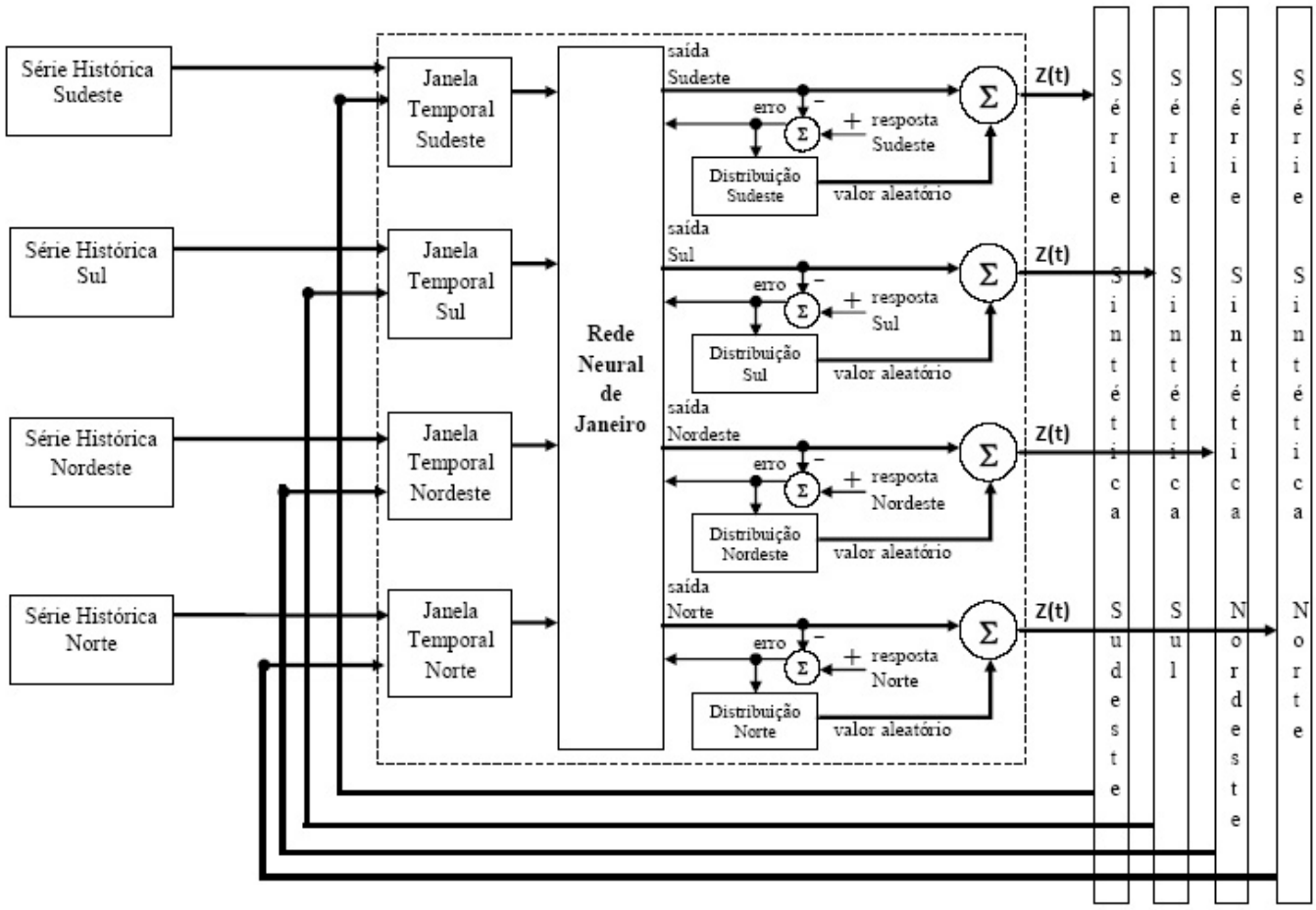

Figura 1: Rede Neural de Janeiro com 4 saídas

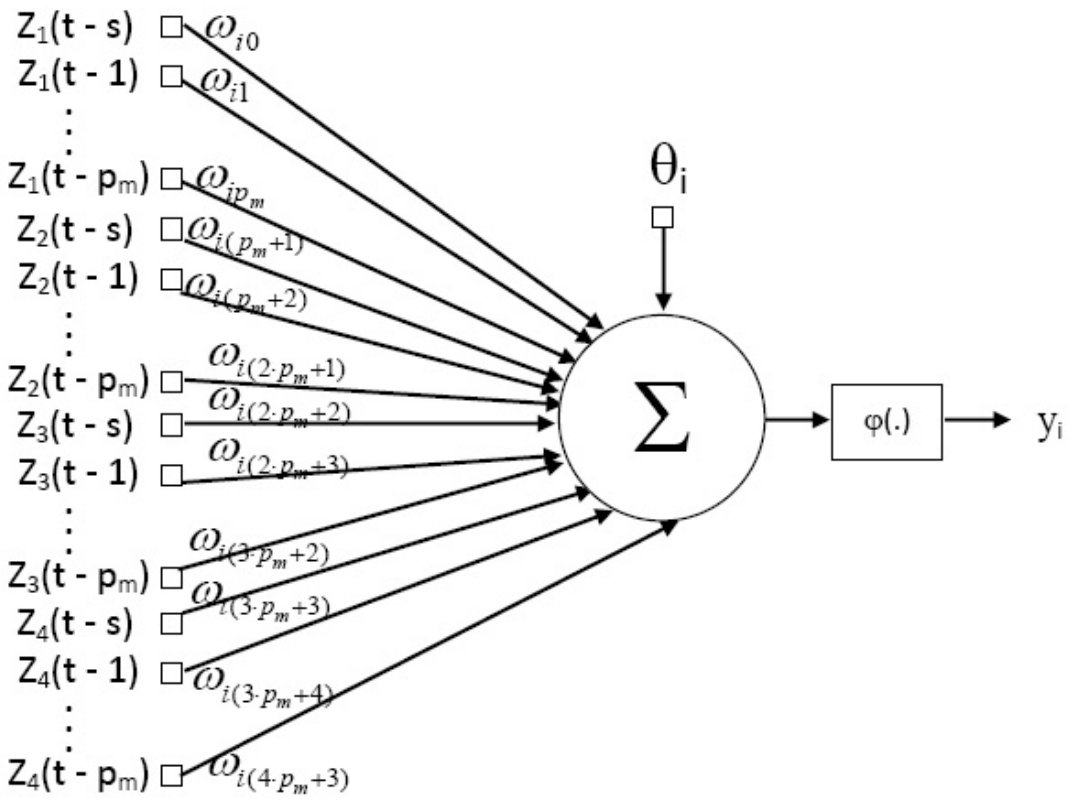

Figura 2: Neurônio da camada oculta da rede neural de 4 saídas

$$\begin{aligned} &y_i = \varphi\left(\sum_{\gamma=1}^{4}\left[\omega_{i,(\gamma-1)\cdot p_m+(\gamma-1)}\cdot Z_\gamma(t-s)+\lambda\right]+\theta_i\right)\\ &onde\\ &\lambda = \left(\sum_{j=1}^{p_m}\omega_{i,(\gamma-1)\cdot p_m+(\gamma-1)+j}\cdot Z_\gamma(t-j)\right)\end{aligned} \tag{2}$$

onde φ é a função de ativação do neurônio i, γ representa o subsistema, p_m é a ordem do mês m, $Z_\gamma(t-j)$ é o valor da série do subsistema γ no índice de tempo $t-j$, $s=12$, θi é o *bias* do neurônio i e $\omega_{i,j}$ é o peso sináptico da ligação entre a entrada j e o neurônio i. Assumindo que $Z_{\gamma(t-s)}$ tem índice $\iota=0$, os outros valores passados da série de ENA do subsistema γ, $Z\gamma(t-\iota)$, tem índice $\iota=1,\ldots,p_m$. Portanto, o índice $j=(\gamma-1)\cdot(p_m+1)+\iota$.

Supondo-se que esta rede neural tem L neurônios na camada escondida, a saída de cada neurônio pertencente à camada de saída é calculada pela equação 3.

$$y_{out}(\gamma) = \varphi_{out}\left(\sum_{i=1}^{L}\omega_{out_\gamma,i}\cdot y_i+\theta_{out}(\gamma)\right) \tag{3}$$

onde $y_{out}(\gamma)$ é a saída do neurônio do subsistema γ, φ_{out} é a função de ativação dos neurônios pertencentes à camada de saída, representa por out, $\omega_{out_\gamma,i}$ é o peso sináptico da conexão entre a entrada i (que corresponde à saída do neurônio i pertencente à camada oculta) e o neurônio out do subsistema γ, onde $i=1,\ldots,L$ e $\theta_{out}(\gamma)$ é o *bias* do neurônio do subsistema γ.

Durante o treinamento desta RNA, para cada padrão k de entrada-saída, um erro entre o valor real da série do subsistema γ, aqui representada por $d(k)$, e a saída prevista para o subsistema γ pela rede neural, é calculado pela equação 4.

$$e_{out}(\gamma)(k) = d(\gamma)(k) - y_{out}(\gamma)(k) \tag{4}$$

Portanto, uma série de resíduos é gerada para cada subsistema γ e uma função de distribuição teórica de probabilidade é definida para cada série. A função de distribuição de probabilidade do subsistema γ é usado para selecionar aleatoriamente um valor a ser acrescentado à saída do neurônio deste subsistema, para compor sua série sintética, de acordo com a equação 5.

$$Z_\gamma(t) = y_{out}(\gamma)+\alpha_\gamma(t) \tag{5}$$

onde $Z_\gamma(t)$ é a serie do subsistema γ gerado no índice de tempo t e $\alpha_\gamma(t)$ é o valor aleatório da função de distribuição de probabilidade ajustada aos resíduos obtido pela diferença entre a saída desejada e a saída da RNA referente ao subsistema γ, representado aqui por $y_{out}(\gamma)$.

Isso é realizado por todas as 12 RNAs do PEN4out. Note que as séries dos subsistemas do SIN são simuladas paralelamente, através da execução seqüencial da equação 5.

3 Estudo de Caso: Geração de Cenários de Afluências Mensais com PEN4out

A ordem p_m de cada janela temporal da rede neural da CE do mês m pode variar de 1 a 11, que corresponde ao número máximo de termos passado permitido a fim de manter a sazonalidade anual da série. De experimentos preliminares, foi observado que as séries mensais precisam de ordens (lags) diferentes. Assim, baseado nesses estudos preliminares, decidiu-se testar no PEN$_{4out}$ quatro tipos de ordem foram testados para cada janela temporal na rede neural de cada CE: *lags* (atrasos) de 3, 6, 9 e 11 meses.

No modelo PEN original, para cada ordem definida, configurações com L = 1 a 20 neurônios na camada escondida foram testados, resultando em um total de 80 configurações para serem avaliadas por cada mês m. No modelo PEN4out, para cada ordem definida, $4 \cdot L$ nós escondidos foram testados, mantendo as 80 configurações para serem avaliadas para cada mês m. Além disso, cada uma dessas 80 configurações de redes neurais foram treinadas 10 vezes com diferentes inicializações de pesos sinápticos.

Para ajustar cada modelo PEN4$out(p, L)$, um histórico de valores da ENA, de 1931 a 2005, para cada subsistema do SIN, foi utilizado neste trabalho. Portanto, para o ajuste das Redes Neurais do modelo PEN4$out(p, L)$ dispomos apenas de 75 dados mensais. Uma vez que o objetivo do modelo é gerar cenários de séries sintéticas e não fazer a previsão das séries de ENA, optou-se por dividir os dados históricos em apenas dois subgrupos: conjuntos de treinamento e de validação. O conjunto de validação foi utilizado para determinar o número ideal de ciclos de formação, no processo de interrupção precoce para evitar *overfitting*, bem como verificar o melhor número de neurônios escondidos. A etapa de validação das redes neurais do PEN4out consiste na geração de uma série com k anos, composta pela concatenação das s saídas das redes neurais em cada ano. Portanto, foram usados os k últimos anos dos dados históricos para compor o conjunto de validação, enquanto todos os anos anteriores definiram o conjunto entrada-saída de treinamento. Neste estudo foi utilizado k igual a 5, que coresponde ao total de anos que o atual planejamento do SIN é realizado. Portanto, o conjunto de validação é composto dos últimos cinco anos dos dados históricos e os dados restantes pertencem ao conjunto de treinamento.

Após o treinamento de todas as configurações de redes neurais, a rede neural do mês m que fornecer o melhor desempenho no conjunto de validação é selecionado para formar a m-ésima CE do PEN4out. Como o conjunto de validação é composto por apenas cinco anos, também foi decidido, a fim de verificar o desempenho das redes neurais treinadas, testar com um número maior de padrões de dados, composto da unificação dos conjuntos de treinamento e validação. A medida de desempenho escolhida para avaliar o desempenho das redes neurais que irão compor as CEs do PEN4out é o percentual de erro médio absoluto (MAPE) [12], que é amplamente usado para validar modelos de séries temporais, calculado pela equação 6.

$$MAPE_\gamma = \frac{1}{x} \cdot \sum_{k=1}^{x} \left| \frac{Z_\gamma(k) - C_\gamma((k-1) \cdot s + m)}{Z_\gamma(k)} \right| \tag{6}$$

onde $\gamma = 1 \ldots 4$ representa um subsistema do SIN, m é o mês, x é o número total de padrões do mês m na série, k é o índice do elemento, $s = 12$ é a quantidade de meses, $Z_\gamma(k)$ é a saída desejada do k-ésimo padrão unificado do mês m do subsistema γ, C_γ é uma série criada com as saídas das redes neurais para o subsistema γ.

Uma vez que a rede neural que compõe a m-ésima CE do PEN4out é selecionada, a série de erros, obtida pela diferença entre a saída da rede neural e a saída do conjunto de treinamento, é usada para ajustar uma função de distribuição de probabilidade teórica para compor esta CE. Portanto, uma CE do modelo PEN4out é formada por uma RNA com quatro saídas e quatro funções de distribuição de probabilidade, conforme ilustrado na Figura 1.

4 Resultados da Simulação

Após o treinamento de todas as configurações de RNA, analisou-se todos os resultados obtidos e a configuração da RNA que proporcionou o melhor desempenho foi escolhida.

Com cada configuração de RNA treinada utilizou-se algumas medidas de desempenho como: MAPE calculado com dados do conjunto de validação, MAPE calculado com dados do conjunto unificado de validação e treinamento, média da série de resíduos obtidos entre a diferença da saída desejada e a saída dada pela rede neural com dados do conjunto de validação e também calculada com dados do conjunto unificado.

A medida de desempenho que ofereceu o melhor resultado para medidas de desempenho foi a média da série de resíduos de validação de cada mês, e portanto foi a medida de desempenho utilizados para escolher a configuração da RNA de cada mês no modelo PEN4out.

A Tabela 1 abaixo apresenta a configuração final de cada rede neural que integra a CE do PEN4out de cada mês. Nesta tabela encontra-se a ordem e o número de neurônios na camada oculta da rede neural escolhida em cada mês.

Com esse modelo final PEN4$out(p, l)$ da Tabela 1, 200 séries sintéticas de 5 anos (60 meses) foram geradas. Com o intuito de analisar a qualidade dessas séries sintéticas geradas, utilizou-se testes estatísticos de aderência [5] para verificar se as séries sintéticas são igualmente prováveis à série histórica de ENA.

Aplicou-se o teste *t* para avaliar se a média das séries sintéticas é estatisticamente igual à média histórica. A hipótese é que a média dos 200 valores de cada mês em cada ano é estatisticamente igual à média histórica do mês correspondente. Em cada

Tabela 1: Configurações da Redes Neurais selecionadas

Mês	1	2	3	4	5	6	7	8	9	10	11	12
Ordem	11	11	3	9	11	9	11	3	3	3	6	3
# Neurônios	64	8	72	28	28	52	4	4	8	64	68	24

teste realizado obtém-se um p-valor e caso esse p-valor esteja acima do nível de significância de 5%, aceita-se essa hipótese. A porcentagem dos 60 p-valores (um para cada mês da série) acima do nível de significância indica o desempenho do modelo PEN para gerar séries sintéticas, indicando o quanto o modelo conseguiu reproduzir o primeiro momento da série histórica.

De maneira análoga foi aplicado o teste de Levene para avaliar se a variância de cada período dos cenários é estatisticamente igual à variância histórica do mês correspondente, e o teste de Kolmogorov-Smirnov (K-S) para verificar se os cenários provêm da mesma distribuição de probabilidade do histórico, indicando que o modelo reproduziu corretamente o comportamento da série histórica.

Como o histórico de ENA foi dividido em dois subgrupos, um para o treinamento e outro para a validação das redes neurais do PEN$4out$, a avaliação mais otimista das séries sintéticas geradas foi realizado com este conjunto de validação, correspondentes aos últimos cinco anos do histórico. A Tabela 2 mostra os resultados obtidos em todos os testes estatísticos.

Tabela 2: Resultados dos testes estatísticos com as séries sintéticas geradas pelo PEN$4out$

Subsystem	Histórico		
	Validação		
Teste	t	Levene	K-S
Sudeste/Centro-Oeste	80%	97%	60%
Sul	60%	94%	62%
Nordeste	49%	87%	39%
Norte	77%	94%	60%

É possível notar que o test t só apresentou aderência um pouco abaixo de 50% em relação às séries de ENA do subsistema Nordeste. O teste de Levene já apresentou valores altos de aderência em todos subsistemas. Já o teste de Kolmogorov-Smirnov (K-S) as aderências para todos os subsistemas não foram muito altas, porém, exceto em relação às séries de ENA do subsistema Nordeste, as séries dos outros subsistemas apresentaram aderência acima de 50%. Isso indica que o comportamento da série sintética gerada tendeu a seguir o comportamento da série histórica. Contudo, nesses experimentos foram utilizadas ordens (lags) iguais para as séries de ENA de todos os subsistemas. Através dos resultados obtidos com o PEN na versão original, acredita-se que isso prejudicou o modelo. Os resultados podem ser melhorados se forem utilizadas janelas de tempo diferenciadas para cada subsistemas, ou seja, se não fixar a mesma quantidade de termos passados iguais para as séries de ENA de todos os subsistemas.

5 Conclusão

O objetivo deste estudo foi elaborar uma nova versão do modelo de processo estocástico baseado em redes neurais, chamado Processo Estocástico Neural (PEN), com o objetivo de incorporar a correlação espacial existente nas séries de Energia Natural Afluente (ENA) de cada subsistemas do Sistema Interligado Nacional (SIN). As ENAs são séries não estacionárias, devido a períodos de cheias e seca do ano, e sazonal, com períodos de 12 meses, que apresentam correlações periódicas temporal e espacial.

Para tratar essa correlação espacial, não presente na versão original do PEN, a rede neural de cada mês recebe as séries de ENA dos 4 subsistema do SIN. Este novo modelo fornece como saída, simultaneamente, as quatro séries sintéticas, uma de cada subsistemas, por isso é chamado de PEN$4out$. O objetivo do presente modelo é a geração de séries temporais sintéticas tão provável como as séries históricas de ENA, cobrindo um período de tempo.

No estudo de caso gerou-se um conjunto de 200 séries sintéticas de 5 anos de ENA para cada subsistema do Sistema Interligado Nacional. Os resultados obtidos mostraram que as séries sintéticas geradas pelo PEN$4out$ apresentam aderências razoáveis em relação às séries históricas analisadas. Contudo, no modelo avaliado neste trabalho, todos os subsistemas utilizam a mesma quantidade de termos passados nas series de todos subsistemas. Acredita-se que o desempenho pode ser melhorado se for permitido ordem diferente para as séries de cada subsistema, como ocorre com o PEN na versão original [6].

6 Agradecimentos

Agradeço a FAPEMIG pelo apoio financeiro concedido para a apresentação desse trabalho.

Referências

[1] "ONS - Operador Nacional do Sistema Elétrico". http://www.ons.com.br.

[2] M. V. F. Pereira and L. M. V. G. Pinto. "Operation planning of large-scale hydroelectric systems". In *Proc. 8th Power System Computation Cont.*, 1984.

[3] I. Luna, R. Ballini and S. Soares. "Técnica de Identificação de Modelos Lineares e Não-Lineares de Séries Temporais". *Revista Controle e Automação*, vol. 17, no. 3, pp. 245–256, 2006.

[4] A. J. G. Abelém. "Redes neurais artificiais na previsão de séries temporais". Dissertação de mestrado, Pontífica Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, Setembro 1994.

[5] D. C. Montgomery and G. C. Runger. *Estatística Aplicada e Probabilidade para Engenheiros*. Editora LTC, Rio de Janeiro, RJ, second edition, 1971.

[6] L. Campos, M. Vellasco and J. Lazo. "A Stochastic Model based on Neural Networks". *2011 International Joint Conference on Neural Networks*, 2011.

[7] K. W. Hipel and A. I. McLeod. *Time Series Modelling of Water Resources and Environmental Systems*. *Elsevier*, Amsterdam, The Netherlands, 1994.

[8] G. Box and G. Jenkins. *Time series analysis: Forecasting and control*. *San Francisco: Holden-Day*, 1970.

[9] S. Haykin. *Redes Neurais: Princípios e Prática*. Bookman, Porto Alegre, RS, 2001.

[10] M. Hagan and M. Menhaj. "Training feedforward networks with the Marquardt algorithm". *IEEE Transactions on Neural Networks,*, novembro 1994.

[11] D. E. Rumelhart, G. E. Hinton and R. J. Williams. "Learning representations by backpropagating errors". pp. 533–536, Nature (London), 323, 1986.

[12] Z. Tang, C. Almeida and P. A. Fishiwick. "Time series forecasting using neural networks vs Box-Jenkins methodology". In *Simulation*, volume 57, pp. 303–310, 1991.