Livro de Minicursos do XIV Congresso Brasileiro de Automática, Natal, 2002.

Introdução ao Uso de Redes Neurais na Modelagem de Sistemas Dinâmicos e Séries Temporais.

Luiz P. Calôba
COPPE & EP – UFRJ
(caloba@ufrj.br)

Resumo: Neste trabalho apresentamos uma introdução ao uso de redes neurais como ferramenta auxiliar na modelagem de sistemas dinâmicos e de séries temporais. Na seção I apresentamos a rede neural *feedforward* que será usada ao longo deste trabalho, e seu treinamento e operação. Na seção II apresentamos brevemente alguns aspectos de sistemas dinâmicos e introduzimos a estrutura NARMA, que será usada para modela-los, discutindo detalhes desta modelagem e do treinamento. Na seção III apresentamos os processos clássicos de análise de séries temporais e como redes neurais podem ser utilizadas como ferramenta auxiliar para tratar resíduos não lineares. Ao término das seções II e III são apresentados exemplos de aplicação. A seção IV apresenta as conclusões e algumas referências bibliográficas.

I. Redes Neurais

Redes Neurais são algoritmos que tentam emular de uma forma muito simplificada a maneira como o cérebro animal processa determinadas informações. São baseadas em processadores elementares chamados neurônios, Fig. 1,

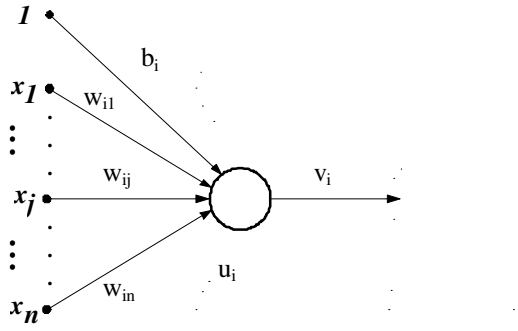


Fig. 1 – Neurônio.

definidos pelas equações

$$u_i = \sum_{j=1}^{n} w_{ij}\, x_j + b_i$$

$$v_i = \begin{cases} u_i & \text{neurônio linear} \\ \text{tgh}\, u_i & \text{neurônio tipo tgh} \end{cases}$$

I.1 – Estrutura da Rede Neural

As redes neurais que utilizaremos neste trabalho são tipo feedforward com duas camadas de neurônios, com múltiplas entradas $z_1, \cdots, z_E$ e uma saída $\tilde{y}$, como apresentado na Fig. 2. O único neurônio da camada de saída tem função de ativação linear. A camada intermediária tem Q neurônios, Q – 1 com função de ativação tipo tgh(.) e um único neurônio com função de ativação linear, cuja função é por em destaque possíveis caminhos lineares entre as entradas e a saída. Todos os neurônios da camada intermediária tem sinapse de polarização, que é dispensável no neurônio de saída. Cada entrada se comunica através de sinapses *w* com todos os neurônios da primeira camada, e a saída de cada neurônio da primeira camada se comunica através de sinapses *t* com o neurônio de saída, na segunda camada. Sem perda de generalidade fazemos t_1 constante, $t_1 = 1$.

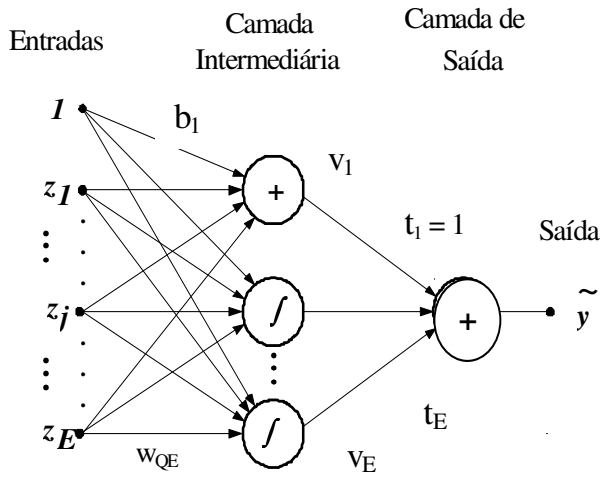


Fig. 2 – Rede neural feedforward usada neste trabalho.

A propagação do sinal é dada por

$$v_1 = \sum_{j=1}^{E} w_{1j}\, z_j + b_1$$

$$v_i = \mathrm{tgh}\left[\sum_{j=1}^{E} w_{ij}\, z_j + b_i\right] \qquad \forall\, i = 2, \cdots, Q$$

$$\tilde{y} = \sum_{j=1}^{Q} v_i\, t_i$$

As variáveis de entrada e saída são escaladas para média nula e amplitude na faixa (-1, 1) antes de serem apresentadas a rede. É possível demonstrar que para um Q suficientemente grande, mas finito, esta rede é um aproximador universal para $\tilde{y} = \varphi(\underline{z})$.

I.2 – Treinamento

O treinamento usual é por épocas. Para a rede da Fig. 2 os acréscimos nas sinapses calculados para descida contra o gradiente (ou retropropagação do erro, regra delta) para cada par entrada-saída $(\underline{z},\ y)$ usando passo de treinamento α são dados por:

$$\varepsilon = y - \tilde{y}$$

$$\Delta t_i = \alpha \varepsilon v_i \quad \forall i = 2, \cdots, Q$$

$$\Delta b_1 = \alpha \varepsilon$$

$$\Delta w_{1j} = \alpha \varepsilon z_j \quad \forall j = 1, \cdots, E$$

$$\Delta b_i = \alpha \varepsilon t_i \left(1 - v_i^2\right) \quad \forall i = 2, \cdots, Q$$

$$\Delta w_{ij} = \alpha \varepsilon t_i \left(1 - v_i^2\right) z_j \quad \forall i = 2, \cdots, Q; \ \forall j = 1, \cdots, E$$

E o acréscimo a ser aplicado após cada época é o valor médio dos acréscimos calculados para cada par entrada-saída pelas equações acima.

Normalmente é necessário controlar o *overtraining* e obter um teste final independente do aprendizado, o que implica em utilizar três conjuntos de pares entrada-saída: o de treinamento, o de validação e o de teste. Cada conjunto deve representar estatisticamente bem a relação entrada-saída da rede.

Mais detalhes específicos do treinamento para modelos de redes dinâmicas e séries temporais serão comentadas posteriormente, nas respectivas seções.

II- Sistemas Dinâmicos

II.1 – Sistemas Discretos no Tempo

São aqueles em que as variáveis são discretas no tempo ou, se contínuas, foram amostradas com período ΔT de modo a torna-las discretas. Para garantir a reconstrução do sinal contínuo original, é necessário que a amostragem seja feita com uma freqüência $f_s = 1/\Delta T$ maior que duas vezes a máxima freqüência significativa envolvida no processo. Usualmente normalizamos $\Delta T = 1$ e $f_s = 1$.

II.2 – Sistemas SISO

São sistemas com uma única entrada x(t) e uma única saída y(t), como na Fig. 3. São os sistemas que consideraremos neste trabalho, sem perda de generalidade: os resultados que apresentaremos podem com facilidade ser estendidos para sistemas com múltiplas entradas e saídas.

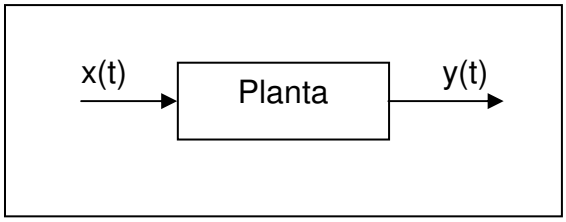


Fig. 3 – Sistema SISO

II.3 – Sistemas Dinâmicos

Sistemas dinâmicos são aqueles em que a saída no instante t depende da entrada no instante t e em T instantes anteriores. Para $\Delta T = 1$,

$$y(t) = \varphi[x(t), x(t-1), \cdots, x(t-T)]$$

II.4 – Modelos para Sistemas Dinâmicos

A equação acima pode ser implementada pela estrutura da Fig. 4 abaixo, composta de uma cadeia de atrasos e um bloco que realiza φ(.)

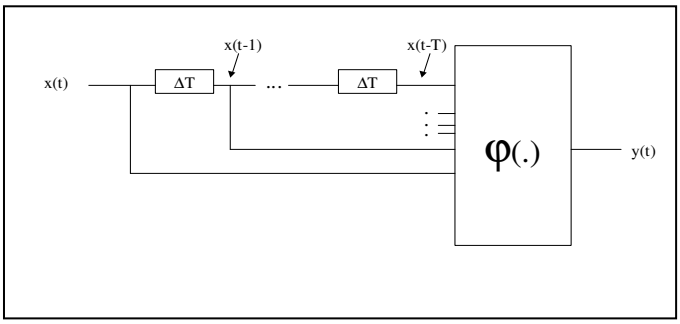


Fig. 4 – Modelo NMA

Se o sistema é linear a equação acima se torna

$$y(t) = \sum_{i=0}^{T} b_i \; x(t - i)$$

e o bloco φ(.) é um mero somador ponderado das variáveis independentes. A estrutura é conhecida como um filtro FIR, *Finite Impulse Response*, em processamento de sinal ou filtro média móvel, MA, em estatística.

Se o sistema é não linear o bloco φ(.) pode ser implementado por uma rede neural feedforward de duas camadas, e a estrutura é conhecida como filtro média móvel não linear, NMA.

A equação do sistema também pode ser escrita com a saída atual como função de entrada atual e de N saídas atrasadas.

$$y(t) = \varphi[x(t), y(t - 1), \cdots, y(t - N)]$$

onde N é a ordem do sistema. A estrutura da Fig. 5 abaixo implementa a equação acima se substituirmos na equação os valores de y(.) por $\tilde{y}(.)$, isto é, em vez das saídas da planta usarmos a aproximação das mesmas calculadas pelo modelo.

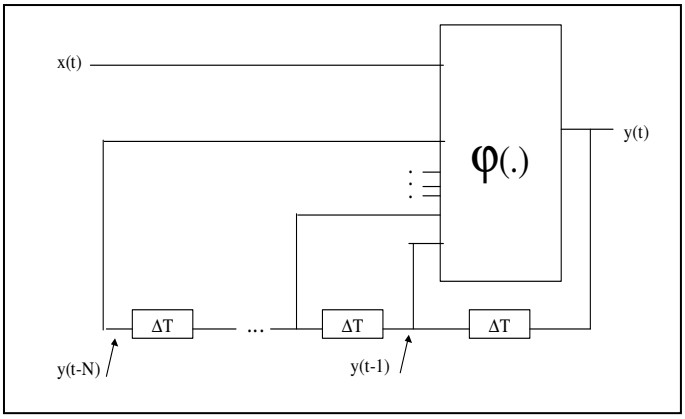


Fig. 5 – Modelo NAR

Se o sistema é linear a função φ(.) é um mero somador ponderado das variáveis independentes, e o modelo é conhecido em estatística como autoregressivo, AR. Se o sistema é não linear φ(.) pode ser implementado por uma rede neural, e o modelo é auto regressivo não linear, NAR, Fig. 5.

Geralmente é mais eficiente associarmos os dois modelos, fazendo a saída função das entradas atual e atrasadas até M intervalo de tempo, e das saída atrasadas até N intervalos de tempo, onde usualmente M ≤ N.

$$y(t) = \varphi[x(t), x(t-1), \cdots, x(t-M), y(t-1), \cdots, t(t-N)]$$

A equação acima pode ser implementada pela estrutura da Fig. 6 abaixo se substituirmos na equação os valores de y(.) por $\tilde{y}(.)$, isto é, em vez das saídas da planta usarmos a aproximação das mesmas calculadas pelo modelo, como no caso anterior.

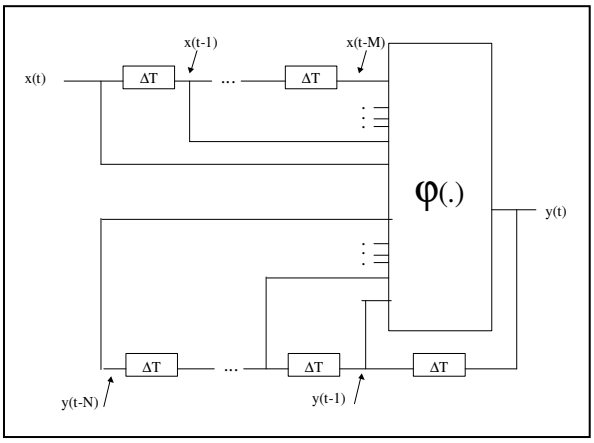


Fig. 6 – Modelo NARMA

Novamente, se o sistema é linear φ(.) é um simples ponderador e a estrutura é conhecida como um filtro IIR, *Infinite Impulse Response*, em processamento de sinais, e filtro ARMA em estatística. Se o sistema é não linear o bloco φ(.) pode ser implementado por uma rede neural e o modelo recebe o nome de ARMA não linear, NARMA.

O modelo NARMA (e suas simplificações NAR e NMA) é o mais utilizado em conjunto com redes neurais, talvez porque pode ser facilmente reinicializado em qualquer instante de tempo, porque seu estado é composto por variáveis do sistema facilmente observáveis, suas entradas e saídas atrasadas.

Existem diversos outros modelos com estados não facilmente observáveis, mas não os consideraremos neste trabalho.

II.5 – Sistemas com Não Linearidades Fracas

Um grande número de plantas dinâmicas reais pode ser representado por um modelo linear sem um erro muito grande. Isto permite imagina-los como sistemas lineares em que uma não linearidade não muito forte foi introduzida. Certamente teríamos interesse em criar um modelo em que as partes linear e não linear que o compõem estivessem separadas, principalmente no caso de não linearidades fracas.

No modelo NARMA genérico a relação entre a saída atual e as entradas e saídas atrasadas é dada pela relação não linear $\varphi(.)$,

$$\tilde{y}(t) = \varphi\left[x(t), x(t-1), \cdots, x(t-M), \tilde{y}(t-1), \cdots, \tilde{y}(t-N)\right]$$

No modelo ARMA, linear, esta relação é

$$\tilde{y}(t) = \sum_{i=0}^{M} b_i \, x(t-i) + \sum_{i=1}^{N} a_i \tilde{y}(t-i)$$

E usando a rede neural da Fig. 2 no modelo NARMA o neurônio linear introduzido na camada intermediária permite escrever

$$\tilde{y}(t) = \sum_{i=0}^{M} w_{1i} \, x(t-i) + \sum_{j=M+1}^{N+M} w_{1i} \tilde{y}(t-j) + \\ + \varphi\left[x(t), x(t-1), \cdots, x(t-M), \tilde{y}(t-1), \cdots, \tilde{y}(t-N)\right]$$

O neurônio linear da camada intermediária da rede modela a parte linear da realimentação, e os (Q-1) neurônios não lineares modelam as não linearidades. Os sub-sistemas linear e não linear estão acoplados apenas pelas respectivas saídas, que se somam antes da realimentação.

II.6 – Dimensão dos Modelos

Um sistema dinâmico real estável tem freqüências naturais $s_i = \sigma_i + jw_i$ com respectivas constantes de tempo $\tau_i = -1/\sigma_i$. Neste trabalho, no caso de sistemas não lineares, chamaremos "freqüências naturais" as freqüências naturais do sistema linearizado nos diversos pontos de operação. O mesmo vale para "singularidades", "constante de tempo", etc. Do ponto de vista prático a saída é independente de entradas atrasadas mais do que quatro vezes a maior constante de tempo do sistema. Isto nos fornece um limite inferior para número de atrasos T no filtro MA ou NMA,

$$T \cdot \Delta T \geq 4 \, \mathrm{Max}\,(\tau_i)$$

Como usualmente as baixas freqüências do sistema são reais,

$$T \cdot \Delta T \geq \frac{2}{\pi f_{min}}$$

onde $f_{min} = 1 / (2\pi \, Max(\tau_i))$ é a mais baixa freqüência natural do sistema, em Hz.

Por outro lado, a freqüência de amostragem f_s deve ser maior que o dobro da maior freqüência operada pelo sistema, f_{max}. Como o sistema é não linear, é razoável considerar esta freqüência f_{max} como pelo menos a maior freqüência natural do sistema ou se for o caso (não provável) alguma freqüência ainda maior inserida na entrada.

$$f_s = \frac{1}{\Delta T} \geq 2 f_{max}$$

Das duas inequações anteriores verificamos que o número de atrasos T necessários no filtro MA é aproximadamente

$$T \geq \frac{f_{max}}{f_{min}}$$

que pode ser razoavelmente grande na prática. Nos modelos ARMA e NARMA, N é a ordem estimada da planta, e $M \leq N$. Usualmente estes modelos são de dimensão muito menor que os MA e NMA, e são mais utilizados na prática.

II.7 – Estabilidade e Precisão dos Modelos

O modelo NMA apresentado na Fig. 4 não inclui realimentação e é portanto estruturalmente estável, e mantém o erro entre a saída da planta e a saída do modelo

$$\varepsilon(t) = y(t) - \tilde{y}(t)$$

com módulo aproximadamente da mesma ordem de grandeza a medida que o tempo evolui.

O modelo NARMA apresentado na Fig. 6 inclui realimentação e é potencialmente instável. Praticamente, se a função φ(.) não for determinada muito precisamente a saída $\tilde{y}(t)$ do modelo diverge da planta y(t) após um certo número de passos de operação.

Isto limita a duração da simulação, e nos obriga a reinicializar a rede após um período de tempo, e retomar a simulação a partir deste ponto. Note que embora inconveniente, no caso do modelo NARMA isto é simples de realizar, porque seu estado real é facilmente determinado a cada instante.

II.8 – Preditores e Operação Série-Paralelo

Há casos, e.g. em controle preditivo e séries temporais, em que desejamos que a saída do modelo no instante atual t aproxime a saída do sistema em um instante futuro t + k, k > 0. Neste caso a equação do modelo NARMA pode ser escrita:

$$\tilde{y}(t+k) = \varphi\left[x(t), \cdots, x(t-M), y(t), \cdots, y(t-N)\right]$$

e implementada por uma estrutura análoga à da Fig. 6, com os valores de y(.) substituídos pelos de $\tilde{y}(.)$ (operação com planta e modelo em paralelo) ou pela estrutura da Fig. 7, (operação com planta e modelo em uma estrutura série-paralelo).

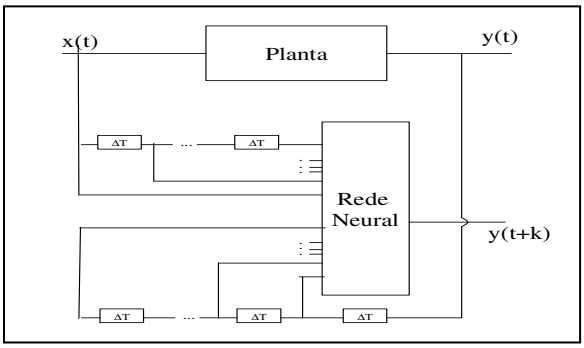


Fig. 7 - Operação em configuração série paralelo.

Na operação em paralelo a informação das saídas atrasadas é fornecida à rede neural pela aproximação feita pela própria rede, eventualmente – provavelmente - realimentando o erro e provocando perda da precisão ao longo do tempo, e instabilidade. A vantagem é que o modelo é completamente independente da planta.

Na operação em série paralelo a informação dos valores de saída atrasados é fornecida à rede pela própria planta, eliminando a realimentação do erro. A estrutura da Fig. 7 não é realimentada, logo é estruturalmente estável. A desvantagem é que exige as informações da planta para operação, limitando as simulações.

II.9 – Detalhes do Treinamento

A rede é treinada fora do modelo, de forma estática. Os pares entrada-saída para o treinamento são obtidos a partir de uma tabela com três colunas, t, x(t) e y(t) que lista a evolução das variáveis de entrada e saída do sistema ao longo do tempo. Cada instante de tempo t maior que N podendo gerar um par entrada-saída ($\underline{e}$, s):

$$\underline{e} = [x(t), x(t-1), ..., x(t-M), y(t-1), ..., y(t-N)]^{t}$$

$$s = y(t)$$

cujos valores numéricos são tirados da tabela com 3 colunas, t, x(t) e y(t).

II.9.1 – Seleção de Pares Entrada-Saída

Uma planta industrial pode permanecer longo tempo no entorno do *set point*, e pouco tempo transicionando de um *set point* para outro. Este processo gerará um grande número de pares entrada-saída no entorno dos *set points* e poucos nas transições, fazendo com que estas não sejam corretamente aprendidas. Usualmente é necessário equilibrar as populações de pares nas diversas regiões ao construir os conjuntos de treinamento e validação. Regiões mal treinadas podem ser detectadas após o treinamento devido aos erros anormalmente grandes que apresentam. Novos pares entrada-saída deverão então ser criados nestas regiões para reforçar a importância dos erros nas mesmas, ou os erros retropropagados ponderados por região.

Um outro detalhe a considerar é que o mapeamento entrada-saída deve obrigatoriamente ser unívoco, o que pode não ser o caso para o modelo inverso de algumas plantas.

II.9.2 – Ordem da Planta

É necessário ter uma idéia a priori da ordem e do tipo da planta para arbitrar N e $M \leq N$. Se escolhermos N pequeno a rede não consegue aprender o mapeamento, e se escolhermos N grande "freqüências naturais" usualmente altas e instáveis são criadas artificialmente. Se não temos uma idéia muito precisa da ordem, uma alternativa é utilizar N e M um pouco maior que o previsto. Após o treinamento examina-se a relevância de cada entrada da rede neural, especialmente aquelas que correspondem aos maiores atrasos. Eliminam-se as entradas de baixa relevância, se houver, e refaz-se o treinamento usando apenas com as entradas mais relevantes.

II.9.3- Singularidades em Altas Freqüências

Plantas reais são normalmente operadas em *set points* escolhidos para maximizar a eficácia no processo, o que geralmente implica em trabalhar em regiões não lineares e com tempo de transição da ordem das maiores "constantes de tempo", para evitar *ringing* ou *overshoot* fortes. Isto é, trabalham com grandes amplitudes de sinal e faixa de freqüências limitada.

Plantas reais podem ter “singularidades” em baixas e altas freqüências, e não linearidades. Os pares entrada-saída devem explorar todas estas características para que o modelo neural aprenda a generalizar a planta verdadeira. As não linearidades e a região de baixas freqüências são normalmente bem exploradas pelos sinais de operação da planta, mas isto muitas vezes não ocorre com a região de altas freqüências.

A solução normalmente permitida é adicionar ao sinal de entrada da planta um pequeno ruído r(t) contendo altas freqüências, usualmente uma onda quadrada com pequena amplitude e *duty cicle* de duração aleatória, que tem um espectro de freqüência mais ou menos constante. Entretanto, como este sinal é muito pequeno comparado ao sinal de entrada, quase não afeta o sinal e o erro na saída da rede, e quase não é percebido pelo algoritmo de treinamento *backpropagation*.

Uma possibilidade é destacar as altas freqüências do sinal de erro. O sinal de erro $\varepsilon(t)$ é separado em dois por filtros com bandas passantes diferentes, o da faixa de baixas freqüência, $\varepsilon_{LF}(t)$, cujas componentes estão na faixa de freqüência do sinal de operação normal da planta, e o da faixa de altas freqüências, $\varepsilon_{HF}(t)$, gerado praticamente pelas altas freqüências componentes do ruído r(t) injetado na planta junto com o sinal de entrada. Um novo sinal de erro $\varepsilon_M(t)$, onde $\varepsilon_{HF}(t)$ é relevante, é gerado e utilizado no algoritmo de treinamento backpropagation, Fig. 8.

$$\varepsilon_M(t) = \varepsilon_{LF}(t) + k\, \varepsilon_{HF}(t)$$

onde $k \geq 1$ inicia grande e é reduzido até 1 ao longo do treinamento.

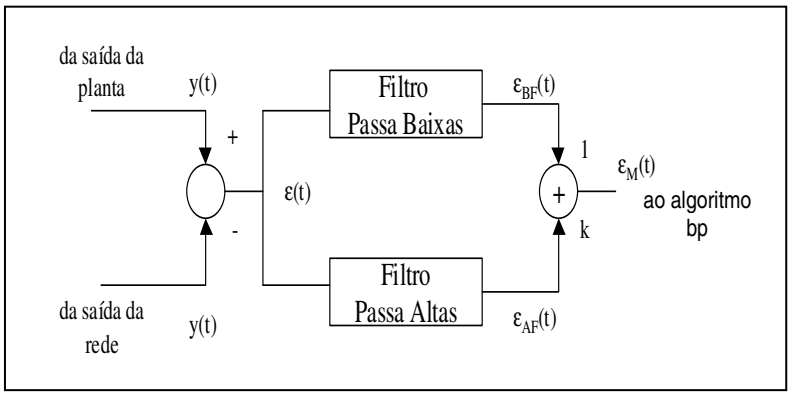


Fig. 8 – Novo sinal de erro

Uma desvantagem deste processo é que os pares entrada-saída devem estar ordenados no tempo, devido aos filtros serem sistemas dinâmicos com estados internos não facilmente determináveis.

II.10 – Exemplos

II.10.1 – Exemplo 1

Considere uma base retangular apoiada em quatro molas e com uma massa desbalanceada sobre ela, Fig. 9.

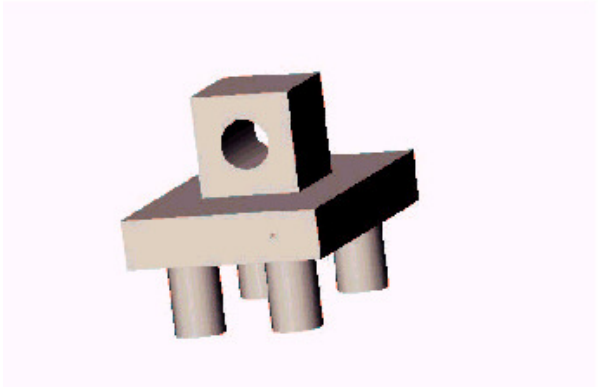
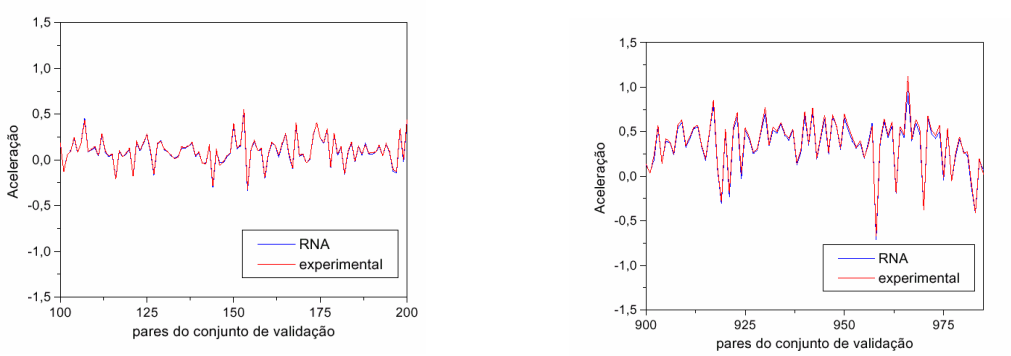


Fig. 9 – Experimento.

Como excitação utilizou-se um martelo de impacto que gera uma entrada x(t) muito breve, praticamente um impulso em nossa escala de tempo, $x(t) = \delta(t)$. A saída y(t) é a resposta de um aceleramento acoplado à placa.

Após o instante inicial t = 0 a entrada é nula, x(t) = 0, e o nosso modelo NARMA reduz-se ao modelo NAR da Fig. 5 em que até mesmo a entrada x(t) é eliminada, porque nula. Após alguns experimentos escolhemos N = 5 e uma rede neural com Q = 3 neurônios na camada intermediária. A rede foi então treinada e testada com os sinais obtidos diretamente do experimento.

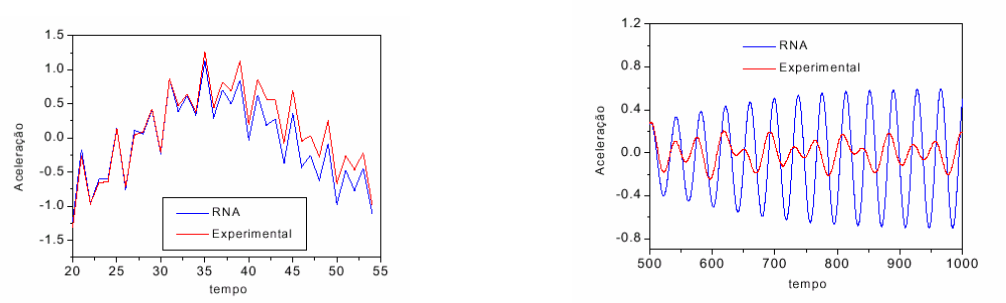
Montada no modo série paralelo, em que as saídas atrasadas são fornecidas pela planta, o modelo representa muito bem a planta em todos os intervalos de tempo, como pode ser visto na Fig. 10.



(a) (b)

Fig. 10 - Resposta da planta e do modelo operando em série paralelo para dois intervalos distintos de tempo, a) 100 < t < 200 e b) 900 < t < 1000

Montado em modo paralelo, em que o modelo se auto-realimenta, a precisão da modelagem somente é boa durante cerca dos 20 primeiros passos da operação, mostrando inclusive tendência a instabilidade para longos tempos de operação, conforme pode ser visto na Fig. 11.



(a) (b)

Fig. 11 - Resposta da planta e do modelo operando em modo paralelo para dois intervalos distintos de tempo a) 20 < t < 55 e b) 500 < t < 1000

Este exemplo destaca o maior problema de uma modelagem do sistema com uma aproximação apenas boa da não linearidade, realizada com dados obtidos diretamente do

experimento. O modelo diverge da planta após alguns passos de simulação, sendo necessário atualizar os estados de tempos em tempos. O que não lhe retira a utilidade.

II.10.2 –Exemplo 2

Considere um sistema dinâmico regido pela seguinte equação a diferença.

$$y(t) = \frac{y(t-1)\, y(t-2)\, y(t-3)\, x(t-2)\, [y(t-3)-1] + x(t-1)}{1 + y(t-3)^2 + y(t-2)^3}$$

Para este exemplo foi utilizado o processo de destacar o erro em altas freqüências descrito na sessão 2.9.3. O sistema foi excitado por uma entrada x(t) composta por um simples sinal senoidal com freqüência baixa, dentro da banda da planta, e amplitude alta, cobrindo toda a faixa dinâmica da saída da planta. A esta senoide foi adicionado um pequeno ruído com amplitude da ordem de 2% da amplitude da senoide.

As condições acima são desconhecidas do modelista do sistema, e foram utilizadas apenas para gerar as tabelas numéricas com as três colunas, o tempo t, a seqüência de entrada x(t) e a seqüência de saída y(t). Desta tabela foram extraídos os pares entrada-saída dos conjuntos de treinamento e validação.

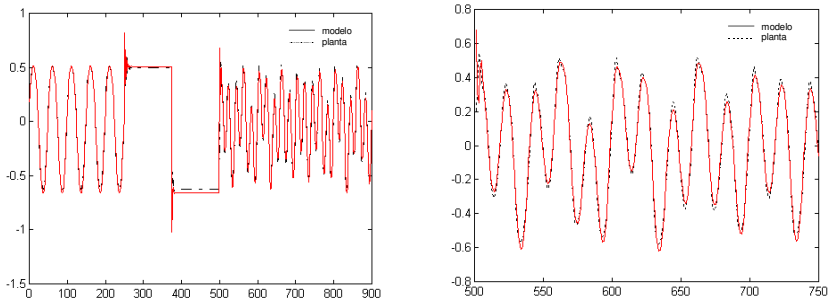
Após algumas observações e simulações iniciais arbitramos para o modelo NARMA que deveria simular o sistema $M = 4$, $N = 5$ atrasos e $Q = 7$ neurônios para a camada intermediária da rede neural.

O treinamento da rede neural foi bastante cuidadoso no tocante à precisão da aproximação. Após dois treinamentos e eliminação das entradas pouco relevantes foram encontrados os valores corretos de $N = 3$ e $M = 2$.

Para dificultar a tarefa do modelo, para o conjunto de teste foi usado como entrada um sinal complexo:

$$x(k) = \begin{cases} \text{sen}(\pi k / 25) & , k < 250 \\ +1 & , 250 \le k < 375 \\ -1 & , 375 \le k < 500 \\ 0{,}3\,\text{sen}(\pi k / 25) + 0{,}1\,\text{sen}(\pi k / 32) + 0{,}6\,\text{sen}(\pi k / 10) & , 500 \le k < 900 \end{cases}$$

A Fig. 12 mostra a resposta da planta e do modelo operando em modo paralelo, isto é, auto-realimentado. O modelo preserva a estabilidade e precisão mesmo após 900 passos de operação, devido à boa qualidade da aproximação provida pela rede neural.



(a) (b)

Fig. 12 - (a) Reposta da planta e do modelo. (b) Detalhe dos últimos passos da operação

É interessante notar que as soluções não são obrigatoriamente únicas. Repetindo o processo de treinamento encontramos M = 1 e N = 2, com resposta surpreendentemente muito semelhante a da Fig. 12. Estados de ordem mais alta podem, em alguns casos, ter pouca relevância na resposta.

III. Séries Temporais

Séries temporais apresentam a evolução de uma ou mais variáveis ao longo do tempo. Mesmo que esta variável seja contínua no tempo, como por exemplo a temperatura em uma localidade, consideraremos que a mesma é amostrada a intervalos de tempo constantes, ΔT, dando lugar a uma série de valores discretos no tempo. A série contínua no tempo $s(t_{contínuo})$ é representada pela série discreta $s(t\,\Delta\,T)$ onde $t = 1, 2, ..., N$. Após a amostragem usualmente normalizamos o intervalo de amostragem $\Delta T = 1$ e consequentemente a freqüência de amostragem para $f_s = 1/\Delta T = 1$. A série discreta passa a ser $s(t)$, $t = 1, ..., N$.

O maior interesse em séries temporais é na previsão, isto é, estando no instante t desejamos prever o valor da variável no instante $s(t + k)$, onde $k > 0$. Mas para poder prever é necessário antes entender, analisar a série.

III.1 – Análise de uma Série Temporal: Decomposição e Transformações

A forma mais prática de analisar uma série temporal é decompo-la em outras séries mais simples. Estas séries mais simples são inicialmente funções determinísticas do tempo. A diferença (erro) entre a recomposição destas séries simples e a série real é uma série residual que normalmente inclui ainda duas outras séries: uma série cujos valores, em cada instante t, dependem de forma complexa, possivelmente não linear, dos valores da série anteriores a t, e uma série de ruído randômico não predizível.

O objetivo da decomposição é, se possível, chegar a uma série residual que seja estacionária no tempo. Uma série é dita estacionária no tempo se todos seus momentos estatísticos são invariantes no tempo. Esta condição é necessária para que os valores passados da série possam ser usadas para caracterizar estatisticamente a série em qualquer tempo, incluir nos tempos futuros em que queremos prever valores. Na prática, limitamo-nos a garantir que os dois primeiros momentos, a média μ e a variância σ^2, sejam invariantes no tempo. Este tipo de série é chamado fracamente estacionária no tempo.

Uma série pode sofrer transformações e decomposições, por exemplo a subtração de outra série ou a multiplicação por uma função do tempo, que a torne estacionária no tempo: Considere a série $s(t) = s_a(t) + f(t)\, s_b(t)$ onde s_a é não estacionária e s_b é estacionária. Fazendo

$$s_1(t) = s(t) - s_a(t) \quad \text{e}$$

$$s_2(t) = s_1(t) / f(t)$$

a série residual $s_2(t)$ é claramente estacionária.

A decomposição pode ser aditiva, multiplicativa ou mista:

$$s = s_1 + s_2 + s_3 + ...$$

$$s = s_1 \,.\, s_2 \,.\, s_3 \,. ...$$

$$s = s_1 + s_2\, [s_3 + s_4] + ...$$

A decomposição vai sendo feita gradualmente, e as funções logaritmo ou exponencial podem ser aplicadas para passar de uma composição multiplicativa para aditiva ou vice-versa. A decomposição aditiva é a mais comum e vamos nos limitar a ela.

O mais importante neste primeiro passo da análise da série talvez seja olhar a representação gráfica da mesma e "ver" as séries que a compõem e que podem ser extraídos. É necessário, mais completo, "ver" a série em três domínios: o primeiro, mais simples e mais importante, é a sua representação gráfica; o segundo é seu correlograma e o terceiro seu espectrograma.

Vamos examinar rapidamente estes dois últimos domínios à seguir.

III.2 - Análise no Domínio do Tempo: Correlação

O coeficiente de correlação de Pearson r(x, y), ou simplesmente correlação, é uma medida da dependência linear entre as variáveis x e y. Considera os pares (x_i, y_i), i = 1, ..., N

$$r(x_i, y_i) = \frac{\sum_{i=1}^{N} (x_i - \mu_x)^2 (y_i - \mu_y)^2}{\left\{ \left[\sum_{i=1}^{N} (x_i - \mu_x)^2 \right] \left[\sum_{i=1}^{N} (y_i - \mu_y)^2 \right] \right\}^{1/2}} =$$

$$= \frac{\frac{1}{N} \sum_{i=1}^{P} (x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y} = \frac{C_{xy}}{\sigma_x \sigma_y}$$

onde μ_x, μ_y, σ_x e σ_x são as médias e desvios padrão de x e y, e C_{xy} é a covariância entre elas.

A correlação r varia no intervalo [– 1 , 1]. Valores grandes do módulo de r, isto é, r próximo de + 1 ou – 1, indicam que existe significativa dependência linear entre as duas variáveis, e valores de r no entorno de zero indicam que não existe esta dependência. A questão é: qual é o limite entre *grande* e *pequeno*, quando existe e quando não existe a dependência ?

Se as duas variáveis x e y forem independentes e randômicas certamente o valor esperado da correlação será 0. Se estudarmos a estatística de r calculada a partir de N pares (x_i, y_i), i = 1, ..., N, verificaremos que a distribuição é normal, a média ou valor esperado é efetivamente nula, e o desvio padrão é $\sigma_r = 1/\sqrt{N}$. Isto significa que, com um nível de confiança de 95% os valores de r de duas variáveis randômicas, sem correlação, estarão entre $\pm 2\sigma_r = \pm 2/\sqrt{N}$ e que praticamente nenhum valor excederá $\pm 3\sigma_r$. 95% é um nível de confiança usualmente adotado. Assim, se ao calcularmos a correlação entre duas variáveis encontramos $|r| > 2/\sqrt{N}$ aceitamos a correlação como existente.

Autocorrelação de uma Série Temporal

A autocorrelação de uma série temporal nada mais é do que a correlação entre o valor atual da série s(t) e o valor atrasado de k unidades de tempo, s(t-k). Se dispusermos de N valores de uma série estacionária no tempo, i = 1, ..., N, poderemos montar N – k pares (s(t), s(t+k)), t = 1, ..., N – k.

$$r(k) = \frac{1}{\sigma_s^2} \frac{1}{N - k} \sum_{t=1}^{N-k} \left(s(t) - \mu_s\right)\left(s(t+k) - \mu_s\right)$$

onde μ_s e σ_s^2 são a média e a variância de s(t). O gráfico de r(k) versus k é chamado de autocorrelograma de s(t).

Correlação Cruzada entre Séries Temporais

A correlação cruzada entre duas séries temporais $s_1(t)$ e $s_2(t)$ estacionárias no tempo mede a correlação entre a variável s_2 tomada no instante atual, $s_2(t)$, e a variável s_1 tomada com um atraso de k unidades de tempo, $s_1(t-k)$. De forma análoga à autocorrelação:

$$r_{s_1 s_2}(k) = \frac{1}{\sigma_{s_1}\sigma_{s_2}} \frac{1}{N - k} \sum_{t=1}^{N-k} \left(s_1(t) - \mu_{s_1}\right)\left(s_2(t+k) - \mu_{s_2}\right)$$

onde $\mu_{s_1}, \mu_{s_2}, \sigma_{s_1}$ e σ_{s_2} são as médias e desvios padrões de $s_1(t)$ e $s_2(t)$. O gráfico de $r_{s1s2}(k)$ versus k é chamado correlograma entre as séries s_1 e s_2.

Correlações somente devem ser calculadas após a extração da tendência da série, ver adiante. A “energia” contida na tendência mascara os detalhes importantes da correlação.

III.3 – Análise no Domínio da Freqüência: Fourier

Uma série s(t), t = 1, 2, ..., N pode ser representada por uma soma de senoides

$$s(t) = a_0 + \sum_{i=1}^{\frac{N}{2}-1} \left[a_i \cos\left(\frac{2\pi}{N} i\, t\right) + b_i \operatorname{sen}\left(\frac{2\pi}{N} i\, t\right) \right] + \frac{a_N}{2} \cos \pi\, t$$

ou

$$s(t) = a_0 + \sum_{i=1}^{\frac{N}{2}-1} R_i \cos\left(\frac{2\pi}{N} it + \theta_i\right) + \frac{a_N}{2} \cos \pi\, t$$

onde $R_i^2 = a_i^2 + b_i^2$ é a "energia" com que a senoide de freqüência $f_i = i/N$ contribui para a série. Os parâmetros a_i e b_i podem ser calculados a partir dos valores de s(t), t = 1,...,N por:

$$a_i = \frac{2}{N} \sum_{t-1}^{N} s(t) \cos \frac{2\pi}{N} i\, t$$

e

$$b_i = \frac{2}{N} \sum_{t-1}^{N} s(t) \operatorname{sen} \frac{2\pi}{N} i\, t$$

mas normalmente é utilizado um algoritmo mais eficiente chamado FFT, Fast Fourier Transform.

O gráfico de $R_i^2(f_i)$ versus $f_i = i/N$ é o espectrograma da série e indica como as diversas freqüências $f_i = 1/N, 2/N, ..., (.5 - 1/N)$ contribuem para a formação da série. Freqüências com contribuições significativamente acima da média merecem tratamento especial por indicarem formas repetitivas no tempo, ou sazonalidades. Se no eixo horizontal em vez da freqüência é apresentado o período N/i da senoide o gráfico é chamado periodograma.

Três observações: (a) séries em tempo contínuo que contem componentes senoidais significativas com freqüências maiores que a metade da freqüência com que foram amostradas geram espectrogramas com erros (principalmente nas altas freqüências) devido

a um fenômeno conhecido como *aliasing*; (b) uma série randômica apresenta um espectrograma com R_i^2 teoricamente constante com a freqüência, assim como as séries das quais foram retiradas as componentes dependentes do tempo e (3) o espectrograma somente deve ser calculado após a extração da tendência da série. Como na correlação, a "energia" contida na tendência mascara os aspectos interessantes do espectograma.

III.4 – Decomposição "Clássica" de Séries Temporais

As séries são geralmente decompostas extraindo-se das mesmas a tendência, a sazonalidade e as componentes senoidais. Mas antes da execução de cada passo de decomposição é recomendável, para tornar o processo numericamente robusto, que a série a ser alterada seja escalada para a faixa de valores de –1 a +1. Algumas vezes pode ser necessário repetir um passo já dado anteriormente, e.g. após a extração de uma componente senoidal pode ser necessário retirar novamente a tendência.

III.4.1 – Tendência

A componente mais fácil de identificar é a tendência tend(t) da série, dada por uma aproximação muito simples da representação gráfica da mesma. É necessário, entretanto, arbitrar a forma da tendência. A mais utilizada é a tendência linear

$$\text{tend}(t) = a_0 + a_1 t, \qquad t = 1, 2, ..., N$$

onde a_0 e a_1 devem ser escolhidos para otimizar a aproximação da série pela sua tendência. A realização da otimização pode por exemplo utilizar o método de mínimos quadrados para calcular a_0 e a_1 que minimizem $F(a_0,a_1)$ abaixo:

$$F(a_0, a_1) = \frac{1}{N} \sum_{t=1}^{N} \left[s(t) - (a_0 + a_1 t) \right]^2$$

Outra alternativa seria utilizar a rede neural abaixo, com sinapses a_0 e a_1 e cujos pares entrada-saída (x,y) são (t, s(t), t = 1, ..., N.

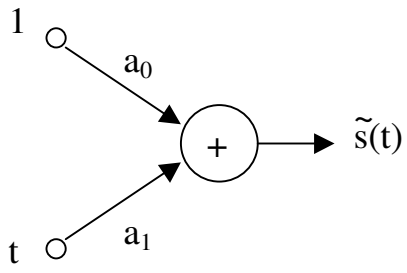


Figura 13 – Rede neural para cálculo da tendência linear

Determinados a_0 e a_1 é gerada a nova série sem tendência $s_1(t)$

$$s_1(t) = s(t) - tend(t)$$

$$= s(t) - (a_0 + a_1 t)$$

Outros tipos mais sofisticados de tendência podem ser utilizados, por exemplo polinomial, logaritmo, exponencial, etc., bem como diferentes funções para diferentes períodos de tempo. Mas temos que lembrar que a tendência extrapola para tempos futuros, e curvas muito rebuscadas podem interpolar bem os pontos do passado e falhar completamente na extrapolação para pontos futuros.

III.4.2 – Sazonalidade

É comum em série temporais termos padrões que se repetem a intervalos regulares de tempo, ou períodos P. Esta característica é chamada de sazonalidade da série. Os períodos da sazonalidade de uma série podem ser determinados pelo conhecimento da fenomenologia da série: por exemplo, espera-se que a venda diária de refrigerantes na praia apresenta forte sazonalidade anual (verão x inverno, férias x não férias) e semanal (fim de semana x dias úteis). Mas a sazonalidade também aparece de forma matemática, caracterizada por uma autocorrelação significativa em atrasos P, 2P, 3P, etc., e por um espectrograma com raias em 1/P, 2/P, 3/P, etc.

Em uma série $s_1(t)$ já sem tendência, determinado o período P, a forma da sazonalidade sz(t) para cada instante i = 1, ..., P do período pode ser calculada como o valor médio dos termos de série neste instante era cada período,

$$sz(i) = \frac{1}{Int(N/P)} \sum_{k=0}^{Int(N/P)-1} s_1(i + kP) \qquad i = 1, ..., P$$

e a série em um tempo t pode ser aproximada pela sua sazonalidade

$$s_1(t) \cong sz\ [Resto\ (t/P)]$$

onde Int(N/P) é o resultado inteiro da divisão N/P, isto é, o número de períodos completos contidos na série, e Resto(t/P) é o resto da divisão t/P, isto é, o instante i da série da sazonalidade que corresponde ao tempo t.

A extração da sazonalidade leva à série $s_2(t)$

$$s_2(t) = s_1(t) - sz\ [Resto\ (t/P)]$$

Existem dois problemas com a sazonalidade. O primeiro é que o período em tempo real da sazonalidade deve ser múltiplo do período de amostragem ΔT; amostragem diária de um fenômeno com uma sazonalidade com período de 1 ano geológico (365,2 dias) somente permite o cálculo direto considerando-se poucos períodos. Este problema pode ser resolvido sofisticando-se um pouco o cálculo.

O segundo problema é que é necessário que a série contenha um número mínimo de períodos, usualmente Int(N/P) > 4, para que a forma da sazonalidade possa ser convenientemente determinada. Isto pode dificultar o cálculo de sazonalidades com longos períodos P, já da ordem do número de pontos N disponíveis na série. Este problema pode ser parcialmente resolvido considerando os ciclos senoidais.

III.4.3 – Ciclos Senoidais

Um caso particular de sazonalidade é o ciclo senoidal, em que a sazonalidade tem a forma de uma senoide com período P. Esta senoide também é caracterizada por uma forma senoidal de período P na autocorrelação da série, e por uma raia de freqüência f = 1/P e amplitude significativamente acima da média em seu espectrograma.

Por serem muito comuns na natureza, forma senoidais são normalmente aceitas como componentes mesmo que a série contenha apenas cerca de 1 ciclo. Por exemplo, manchas solares e séries econômicas apresentam ciclos aproximadamente senoidais com períodos de aproximadamente 10 anos; para séries econômicas este período é usualmente bastante longo quando comparado com o comprimento da série.

A análise de Fourier da série, usualmente através da FFT, nos dá os parâmetros a, b e f do ciclo senoidal da série $s_2(t)$,

$$cs(t) = a \cos(2 \pi f t) + b \operatorname{sen}(2 \pi f t)$$

A série $s_3(t)$, sem esta componente senoidal, é dada por

$$s_3(t) = s_2(t) - cs(t)$$

Dois pontos necessitam atenção. Se o número N de pontos da série $s_2(t)$ é muito pequeno a resolução em freqüência da FFT não é boa, $\Delta f = 1/N$. Se a freqüência da componente senoidal real f estiver entre dois valores de freqüência da FFT, f_j e $f_{j+1} = f_j + 1/N$, o algoritmo fará a aproximação e poderá indicar as 2 (ou mais) freqüências vizinhas de f, f_j e f_{j+1}, como significantes, o que não corresponde a realidade.

Outro problema é que para a série discreta poder reconstruir a série a tempo contínuo corretamente é necessário que esta última não contenha componentes significativas com freqüência maior que a metade da freqüência de amostragem. Esta condição é violada para muitas séries temporais, porque o período de amostragem é fixado em função da fenomenologia da série, e.g. temperatura média **diária**, lucros **mensais**, etc., e os valores da série são adquiridos sem filtragem de freqüências, para não altera-los. A

conseqüência é uma distorção de espectro (aliasing), principalmente em altas freqüências, que falseia os parâmetros a e b fornecidos pela FFT.

O meio mais simples de contornar os dois problemas acima é calcular os parâmetros a, b e f corretos de forma a otimizar no domínio do tempo a aproximação da série pela sua componente senoidal, isto é, calcular a, b e f para minimizar F(a,b,f) abaixo

$$F(a, b, f) = \frac{1}{N} \sum_{t=1}^{N} \left[s_2(t) - (a \cos 2\pi f\ t + b \operatorname{sen} 2\pi f\ t) \right]^2$$

Um bom ponto de partida para a otimização são os valores de a, b e f dados pela FFT.

III.4.4 – Irregularidades

Este item deve ser analisado e corrigido **antes** de iniciarmos a decomposição da série, mas entenderemos melhor um comportamento "irregular" agora que conhecemos os comportamentos "regulares" de uma série, como tendência, sazonalidade, etc.

Uma irregularidade é uma mudança brusca e não previsível de uma das componentes da série, e normalmente corresponde a uma alteração brusca e não previsível no sistema que gera a série. Na Figura 14 está apresentada a série de arrecadação do ICMS do Estado de São Paulo e indicadas no tempo várias diferentes ações governamentais que alteraram de diferentes modos as componentes da série. Estas alterações necessitam ser compensadas de algum modo antes da série ser decomposta.

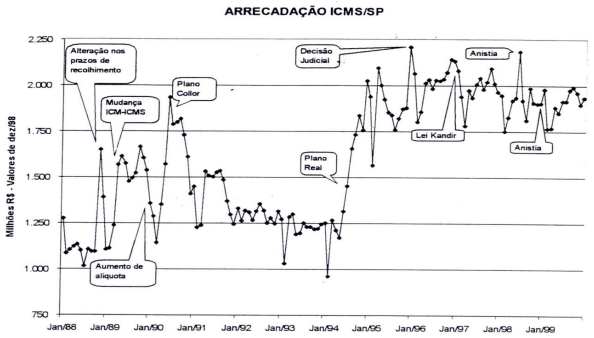


Figura 14 – Arrecadação do ICMS em São Paulo

III.5 – Preditor Não Linear

A série $s_3(t)$ não tem mais componentes que sejam facilmente determinadas como função de tempo, mas seus valores podem ainda ter dependência de valores passados. Isto pode ser verificado através do autocorrelograma de $s_3(t)$. Se forem encontradas correlações significativa de valores atrasados com os atuais, valores futuros podem ser expressos em função dos valores passados. Em princípio até uma simples combinação linear dos valores passados pode ser usada, mas é mais conveniente permitir que o valor futuro seja previsto como uma função não linear dos valores passados, implementada por uma rede neural, atuando como um preditor, como descrito na seção 2.7.

Considere a série $s_3(t)$ que não contém componentes dependentes do tempo. Usualmente esta série residual apresentará uma autocorrelação decrescente com o atraso, isto é, correlação mais significativa para os menores atrasos, embora isto não seja uma regra. Se queremos um previsor para k passos no futuro, isto é, no tempo t queremos prever o valor de $s_3(t + k)$, devemos examinar a autocorrelação de $s_3(t)$ e selecionar atrasos com correlação significativa, mas maiores ou iguais a k: estes são os valores que estarão disponíveis no instante t do cálculo.

Se k é pequeno estamos fazendo um previsor para curto prazo. Provavelmente encontraremos atrasos $t < k$ em que ainda existe correlação significativa, e o preditor será

alimentado com valores **reais** atrasados de $s_3(t)$. Isto corresponde a operar o modelo NMA do preditor em modo série-paralelo, normalmente preciso e estável.

Para previsões de mais longa duração, pode acontecer que não mais exista correlação entre os valores de $s_3(t)$ para longos atrasos, $t \geq k$. É possível neste caso fazer um preditor para atrasos mais curtos, e realimenta-lo com os valores atrasados **previstos** por ele, mas não com os valores reais. Isto corresponde a operar o preditor como um modelo NMA em modo paralelo, auto-realimentado, e portanto sujeito a erros crescentes no tempo e até mesmo instabilidade. O treinamento da rede neural necessita ser muito mais cuidadoso e preciso neste caso.

Finalmente, é possível que exista correlação significativa para atrasos longos, $t > k$ e para atrasos curtos, $t < k$. Para os atrasos longos usamos os dados reais, e para atrasos curtos dados previstos. Isto corresponde a operar o preditor como um modelo NMA parte paralelo e parte série-paralelo.

III.5.1 – Outras Entradas

Outras variáveis podem também ser usadas como entrada da rede neural. Alguma pequena dependência do tempo pode ainda persistir, então t deve ser considerado como entrada. É comum que a distorção não linear seja função do nível de operação da série, que pode ser estimada por

$$\tilde{s}(t) = \text{tend}(t) + \text{sz}(t) + \text{cs}(t)$$

A utilização ou não destas variáveis t, $\tilde{s}(t)$, etc. como entradas da rede neural será aceita se elas apresentarem correlação significativa com $s_3(t)$.

III.5.2 – Séries Auxiliares

Além da série alvo s(t) algumas vezes dispomos também de séries auxiliares sa(t) que mantém correlação com a série alvo e que podem ser utilizadas no preditor da série residual $s_3(t)$. Como $s_3(t)$ é uma série de resíduos de s(t), é bastante mais provável que a

série de resíduos $sa_3(t)$ da série auxiliar sa(t) tenha maior correlação com $s_3(t)$ do que a própria série auxiliar sa(t). Assim, cada série auxiliar sa(t) deve ser decomposta da mesma forma que a série principal, e ter sua tendência, sazonalidade e componentes senoidais extraídas, gerando uma série de resíduos $sa_3(t)$. A correlação cruzada entre as séries de resíduos $s_3(t)$ e $sa_3(t)$ informará se esta última tem contribuição a dar ao previsor, e com que atrasos. Caso as correlações sejam significativas, valores atrasados de $sa_3(t)$ também serão utilizados como entradas da rede neural previsor de $s_3(t)$.

III.6 – Aproximação Final

Todas as operações de escalamento e extração de componentes tem que ser revertidas na reconstrução da série original. Com a decomposição que realizamos a série s(t) no instante t + k é representada por:

$$s(t+k) = \tilde{s}(t+k) + \varepsilon(t+k)$$

onde

$$\tilde{s}(t+k) = tend(t+k) + sz(t+k) + cs(t+k) + \tilde{s}_3(t+k)$$

$\tilde{s}(t+k)$ é a previsão de s(t+k). Tend(t), sz(t) e cs(t) são funções analíticas do tempo e $\tilde{s}_3(t+k)$ é a saída do preditor neural de $s_3(t+k)$, que utiliza apenas as informações disponíveis até o tempo atual t. Se o processo foi bem sucedido a série de erros $\varepsilon(t)$ é uma série randômica e não correlata, um ruído branco.

III.7 – Séries Temporais e Sistemas Dinâmicos

Uma série temporal pode ser aproximada pela soma das saídas de dois sistemas. O primeiro inclui a tendência e a sazonalidade da série, e é um sistema cuja saída é uma função determinística do tempo.

O segundo é um preditor neural, representado por um sistema dinâmico excitado por um impulso no instante t = 0, e é nele que concentraremos nossa atenção. Neste caso o modelo será do tipo NMA, Figura 5, onde até a entrada x(t) pode ser eliminada para $t > 0$, porque é nula. Mas o modelo pode também receber outras entradas relevantes, se

existirem, inclusive entradas atrasadas de séries auxiliares, transformando-se em um modelo NARMA.

Se a série é estacionária no tempo e sem tendência o sistema que modela a série é não dissipativo, uma vez que a variância da série, ou “potência” da saída do sistema, é constante. Isto justifica buscar padrões repetitivos no tempo.

Previsões a curto prazo podem ser feitas baseadas em um modelo operado em série paralelo, naturalmente estável e preciso. Previsões a longo prazo podem ser feitas utilizando um modelo operado ao menos parcialmente em paralelo, que pode levar a erros excessivos em alguns casos. É possível também utilizar modos mistos de operação, em que atrasos curtos são gerados por preditores, e atrasos longos são atrasos reais fornecidos pela própria série.

III.8 - Exemplos

III.8.1 - Exemplo 1

A Fig. 15 mostra a produção mensal de cerveja na Austrália de 1957 a 1994, para a qual desejamos fazer uma previsão. Nota-se imediatamente que em 1976 a série sofre uma mudança brusca em sua tendência. Como o número de dados após 1976 é suficientemente grande, utilizaremos apenas este período.

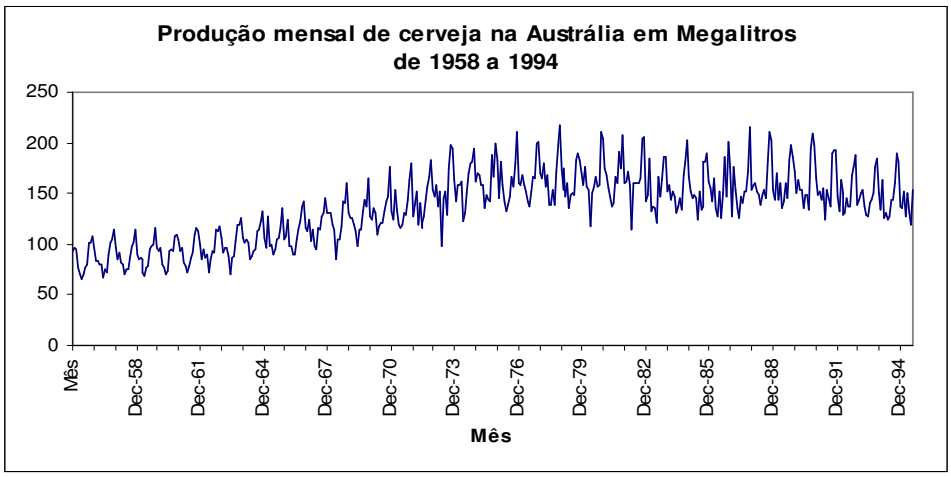


Fig. 15 – Série completa

Reservamos os últimos 20 meses para o conjunto de teste. O período de 1976 a 1992 e a respectiva tendência linear estão mostrados na Fig. 16, e a série sem tendência na Fig. 17.

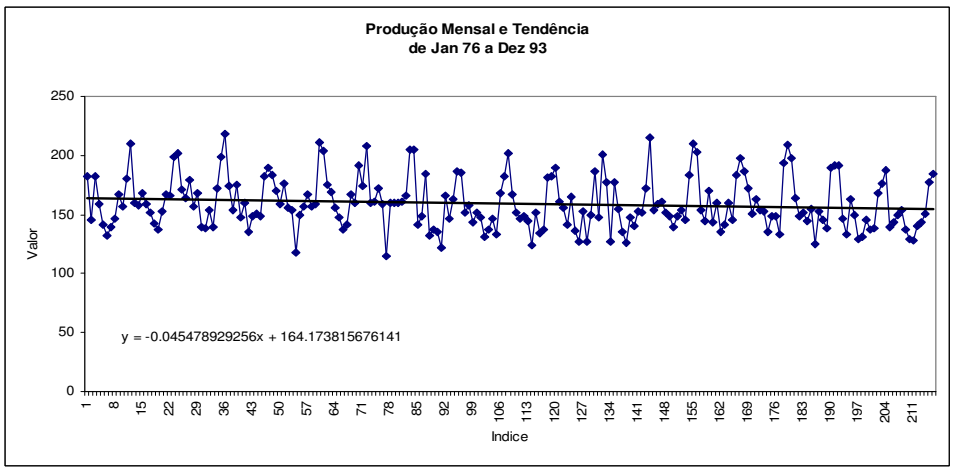


Fig. 16 – Série truncada e tendência linear

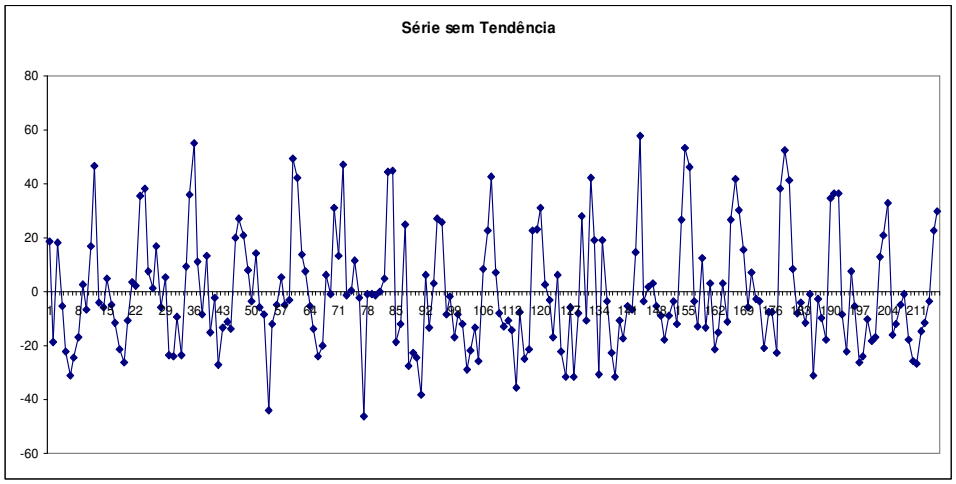


Fig. 17 – Série sem tendência

Claro que em uma série de produção de cerveja esperamos uma sazonalidade anual. A forma da sazonalidade é apresentada na Fig. 18, e a série sem sazonalidade na Fig. 19.

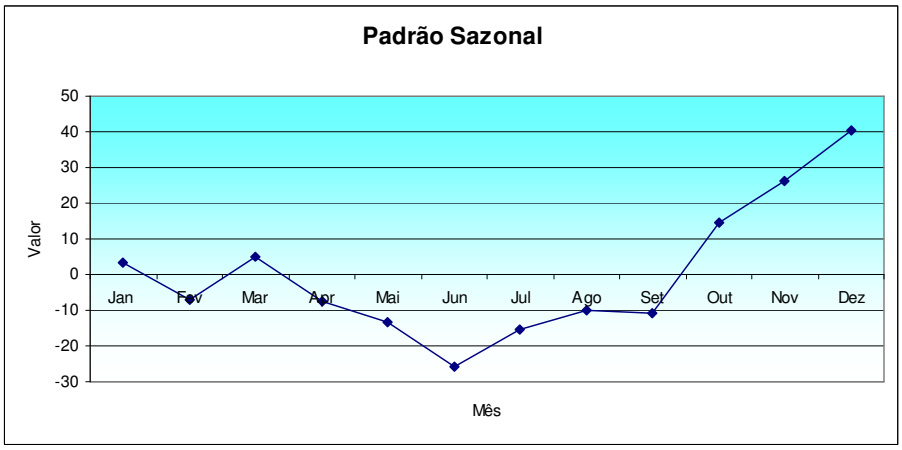


Fig. 18 – Sazonalidade aditiva anual

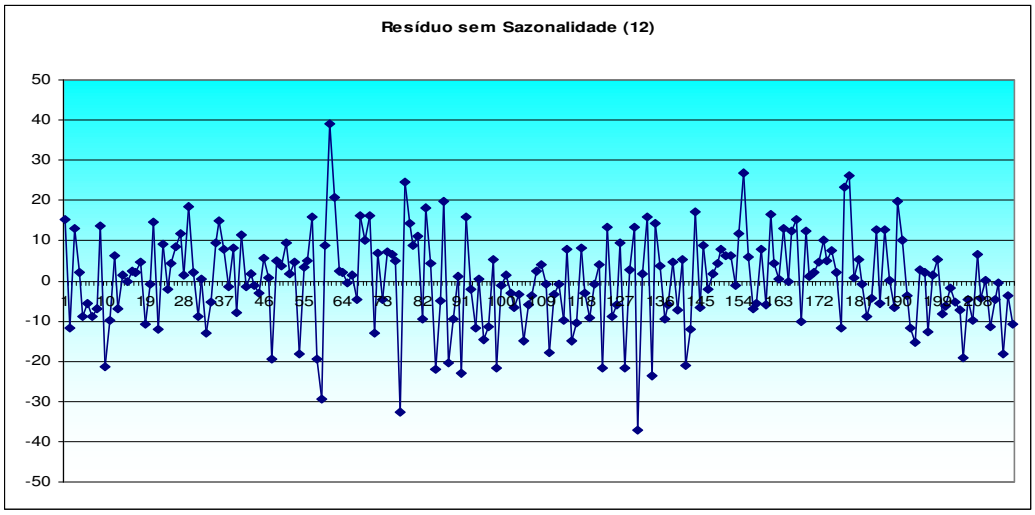


Fig. 19 – Série sem sazonalidade

A Fig. 20 apresenta o espectrograma obtido por Fast Fourier Transform, FFT, da série da Fig. 19, sem tendência nem sazonalidade anual. Podemos notar duas freqüências claramente dominantes. A freqüência mais baixa pode ser extraída com os dados fornecidos pelo programa de cálculo da FFT, mas a mais alta obrigou a realização de uma otimização no domínio do tempo, devido a superposição de espectros (aliasing).

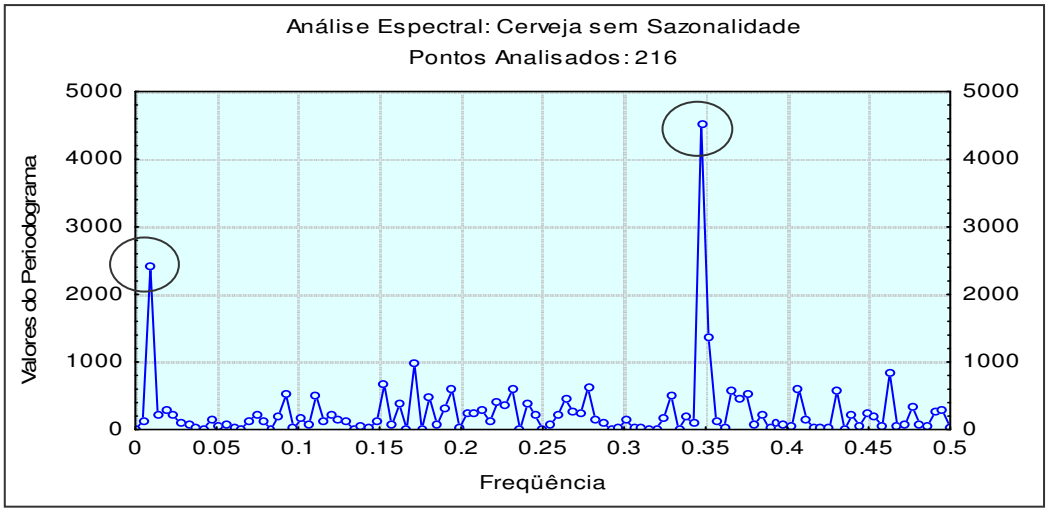


Fig. 20 – Espectrograma da série da Fig. 19

A Fig. 21 apresenta a série residual com as duas componentes senoidais dominantes já excluídas, e a Fig. 22 o espectrograma da série da Fig. 21, mostrando claramente a

eliminação das raias dominantes da Fig. 20. Na Fig. 22 não há mais nenhuma raia dominante que mereça ser extraída.

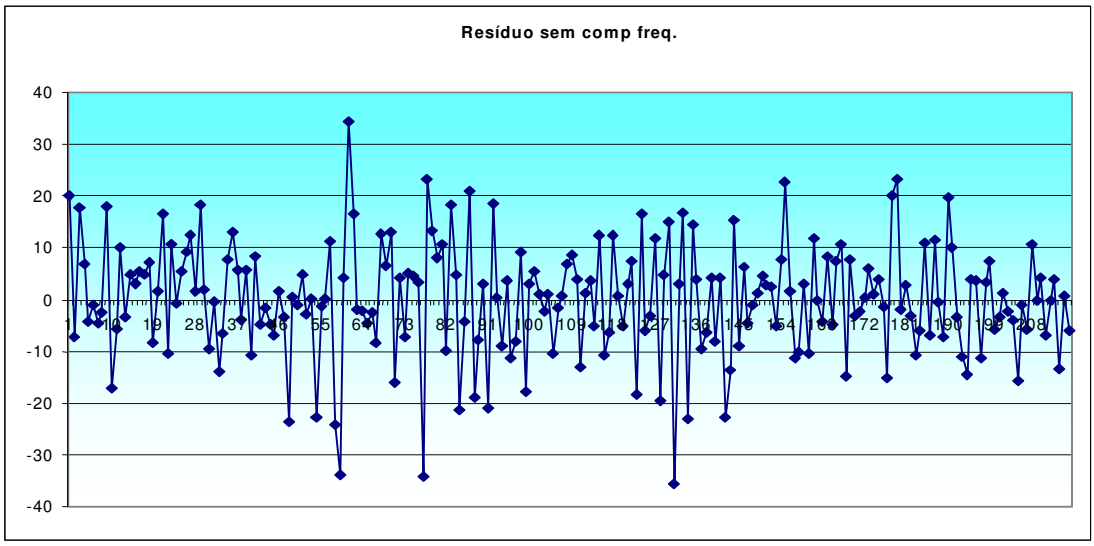


Fig. 21 – Série residual

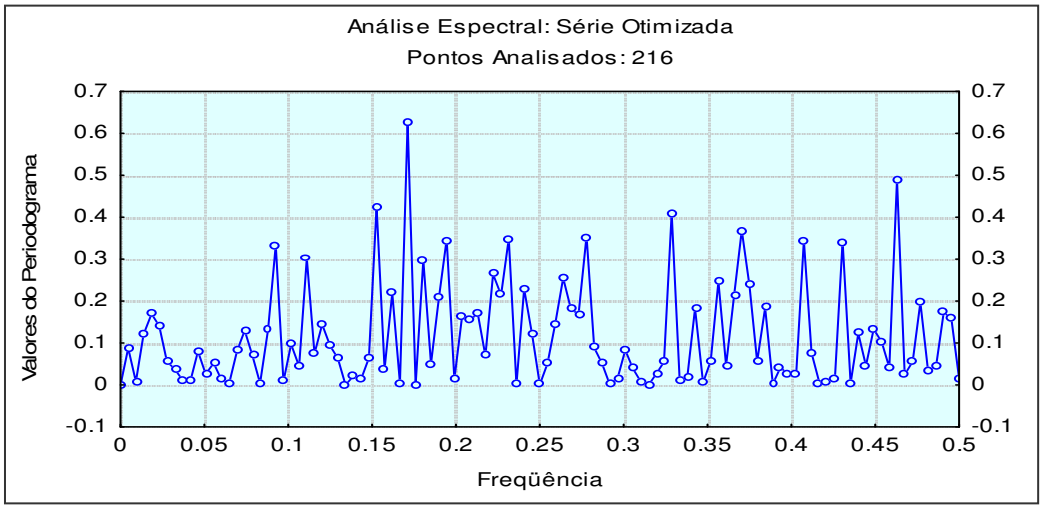


Fig. 22 – Espectrograma da série da Fig. 21

A Fig. 23 mostra a autocorrelação e os limites da confiança de 95% da série da Fig. 21. Um pequeno número de atrasos apresenta correlação significativa. Utilizamos como entradas da rede neural os valores da série com os atrasos de 11, 38 e 50 meses, que mostram claramente confiança maior que 95 %..

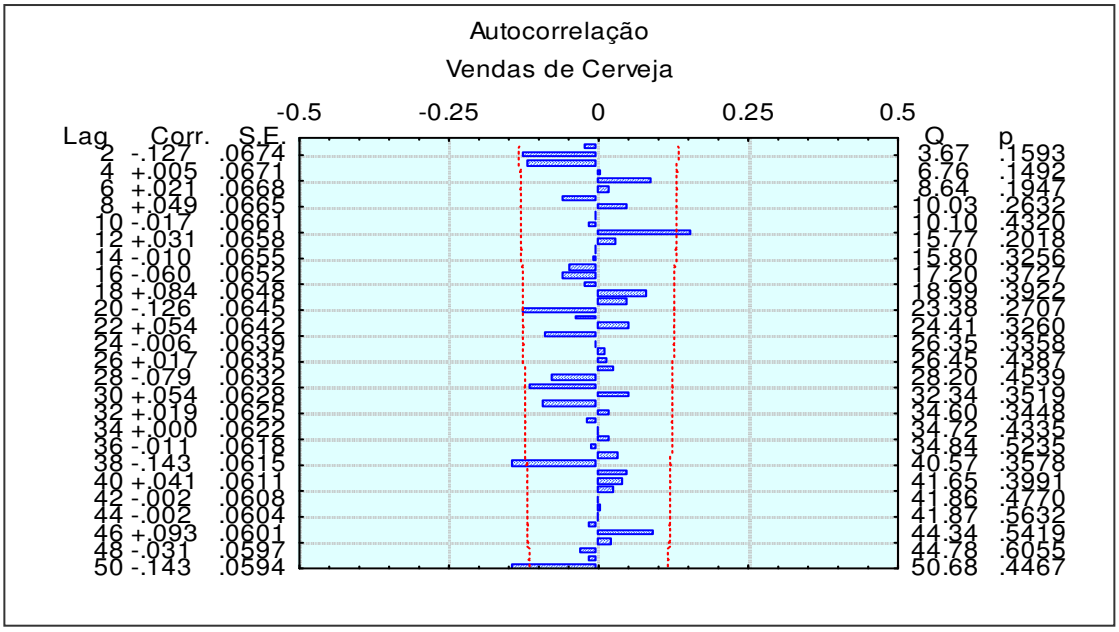


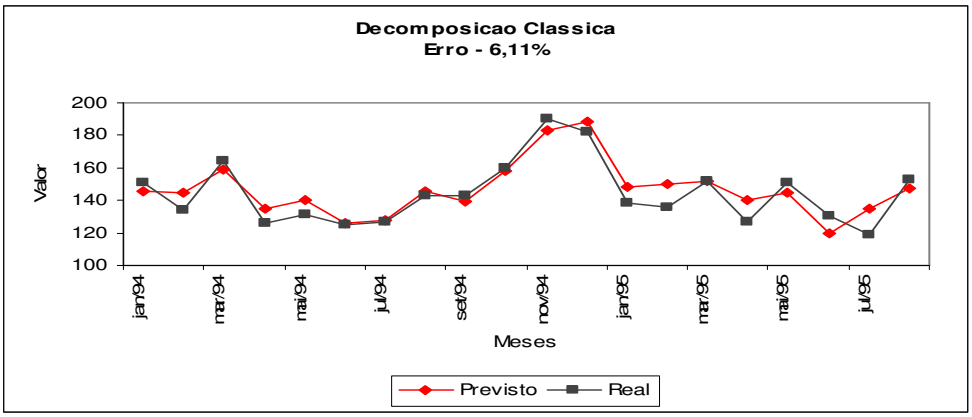
Fig. 23 – Autocorrelação da série residual da Fig. 21.

Experimentalmente chegou-se a uma rede com 6 neurônios na camada intermediária para previsão do valor da série residual com até 11 meses a frente (11 meses é o menor atraso utilizado na entrada da rede).

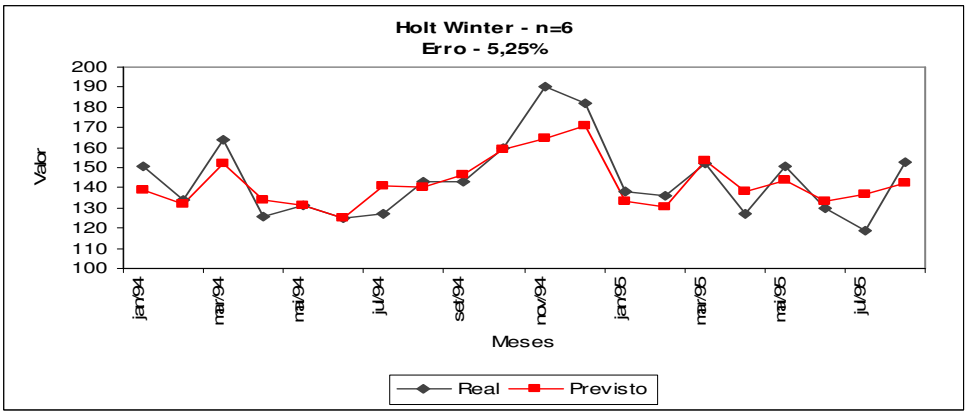
A Fig. 24.a mostra para o período de teste a saída real e a prevista com a decomposição clássica: apenas tendência, sazonalidade e ciclos senoidais. A Fig. 24.b mostra as mesmas curvas usando o modelo de Holt Winters, de amortecimento exponencial triplo, aqui aplicado para fins de comparação. Finalmente, a Fig. 24.c mostra os resultados obtidos com a rede neural usando a decomposição clássica como pré-processamento.

A Tabela abaixo mostra os erros relativos absolutos médios, MAPE, para os três métodos, durante os primeiros 12 meses e durante os 20 meses do conjunto de teste. Em todos os casos a rede neural com pré-processamento apresenta desempenho superior.

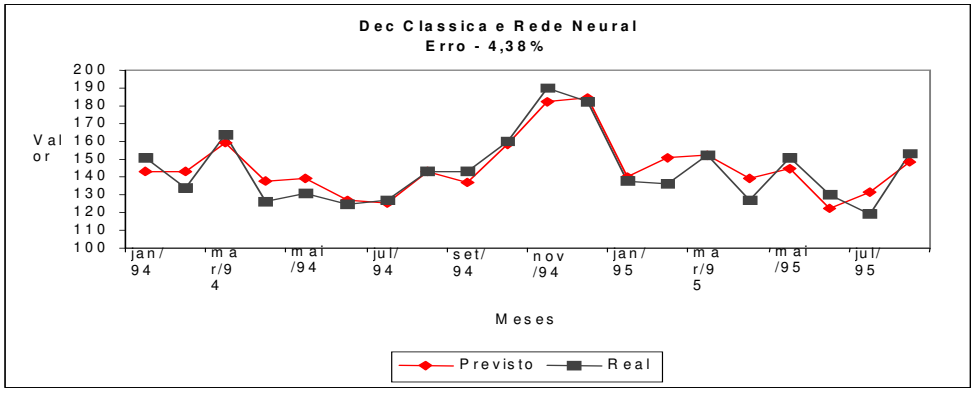
	Decomp.	**Holt Winter**	**RN**
1º Ano	4.71%	4.81%	3.68%
20 Meses	6.11%	5.25%	4.35%



(a)



(b)



(c)

Fig. 24 – Período de Teste, série real e previsão com modelos: (a) decomposição clássica, (b) Holt Winters e (c) Rede Neural com Pré-processamento.

III.8.2 - Exemplo 2:

Este exemplo utiliza Redes Neurais para prever as vazões diárias futuras de um rio em determinado posto fluviométrico a partir das vazões em dias precedentes no mesmo posto, em postos em afluentes situados a montante, e do índice pluviométrico da região. Foi prevista a vazão no posto São Félix, da bacia hidrológica do Alto Tocantins. A montante de São Félix encontram-se os seguintes postos: Colônia dos Americanos, Porto Rio Bagagem, Ceres, Jaraguá, Ponte Quebra-Linha, Tocantinzinho e Porto Uruaçu.

Os dados de entrada foram coletados entre 01/01/1974 e 01/01/1980, perfazendo 2192 dias ininterruptos de coleta. O período correspondente aos anos de 1974 a 1978, foi utilizado para a formação dos conjuntos de treinamento e validação, que guardam entre si uma relação de tamanho de 4/1. Os dados correspondentes ao ano de 1979, foram usados como conjunto de teste da rede, isto é, utilizados para simular a rede em condições reais de operação, depois desta ter sido treinada.

A série alvo, a vazão em São Félix, esta apresentada na Fig. 25, e as séries auxiliares de vazão estão na Fig. 26.

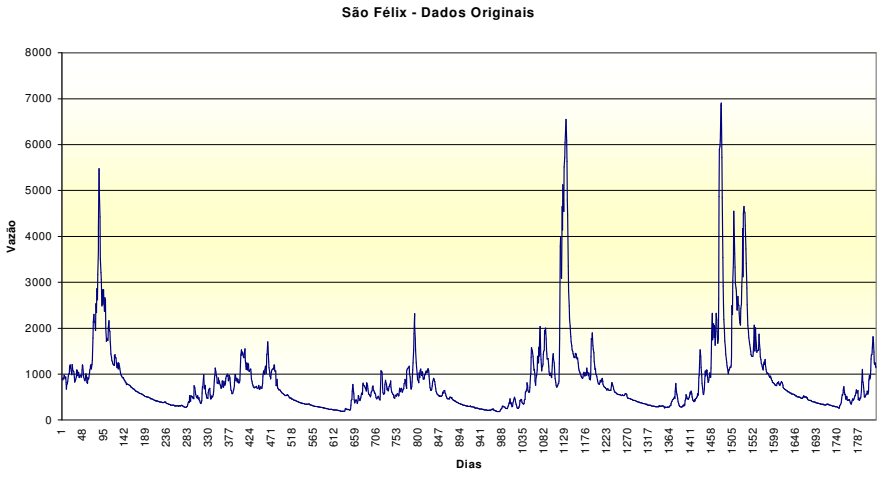


Fig. 25 – Série alvo: Vazão em São Félix.

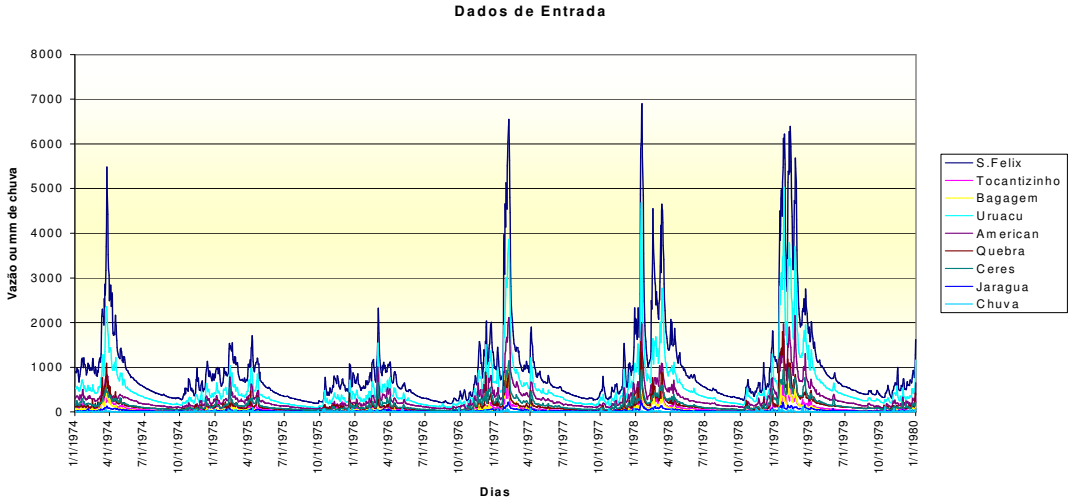


Fig. 26 – Séries Auxiliares: Vazões em todos os postos.

Pré-processamento da Série Alvo:

Escala:

Como havia grande variação de valores nos dados de entrada foi aplicada à série a função logarítmo natural ln (.) com a intenção de suaviza-la e concentrar os dados em uma faixa dinâmica menor, resultando na série da Fig. 27.

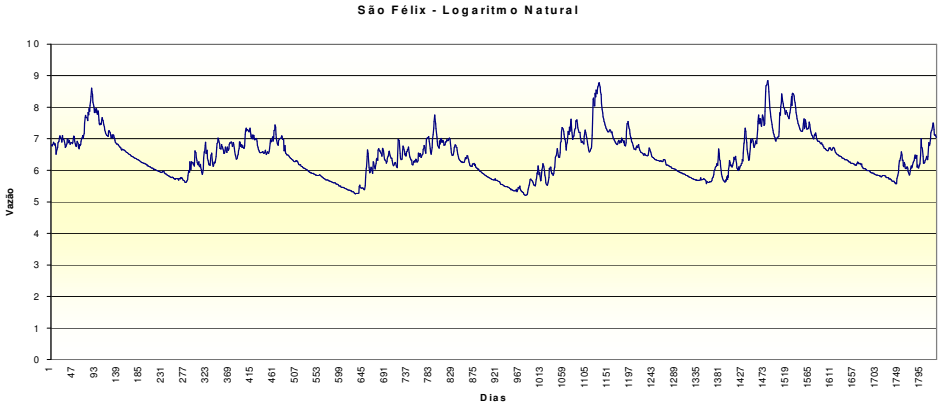


Fig. 27 – Vazão em São Félix, escala logarítmica.

Retirada da Tendência

A tendência linear foi calculada e extraída, e a série resultante esta apresentada na Fig. 28.

$$tend\ (t) = 10{,}23.10^{-6}\ x + 6{,}4579$$

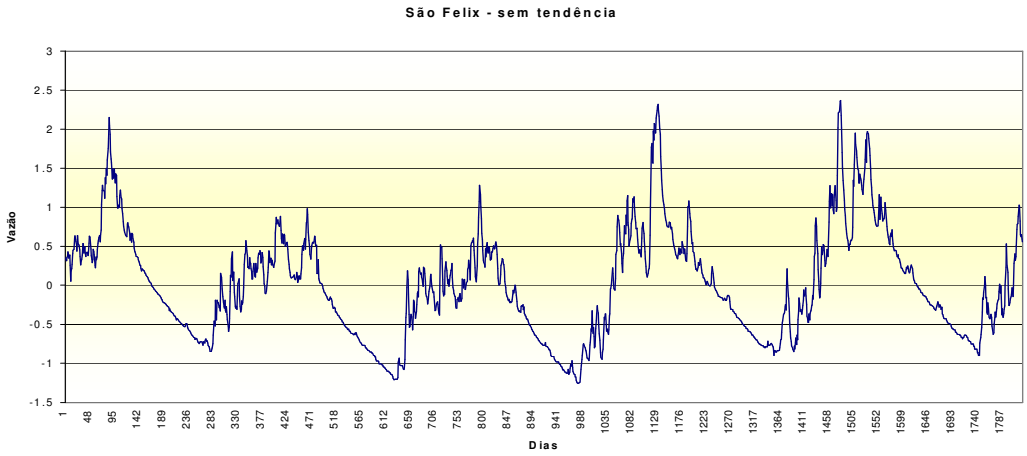


Fig. 28 – Série sem tendência.

Retirada da Sazonalidade

A Fig. 29 apresenta o espectrograma da série sem tendência da Fig. 28, onde é observada uma forte componente com período de 365.2 dias, o período de um ano. Para retirar esta sazonalidade foi utilizado o método das médias móveis, com período de 365 dias. A série com a sazonalidade anual extraída é apresentada na Fig. 30.

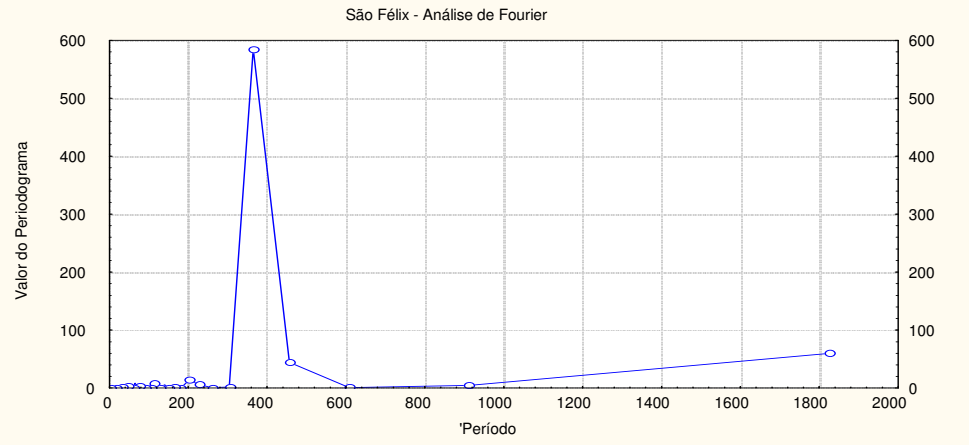


Fig. 29 – Periodograma da série sem tendência.

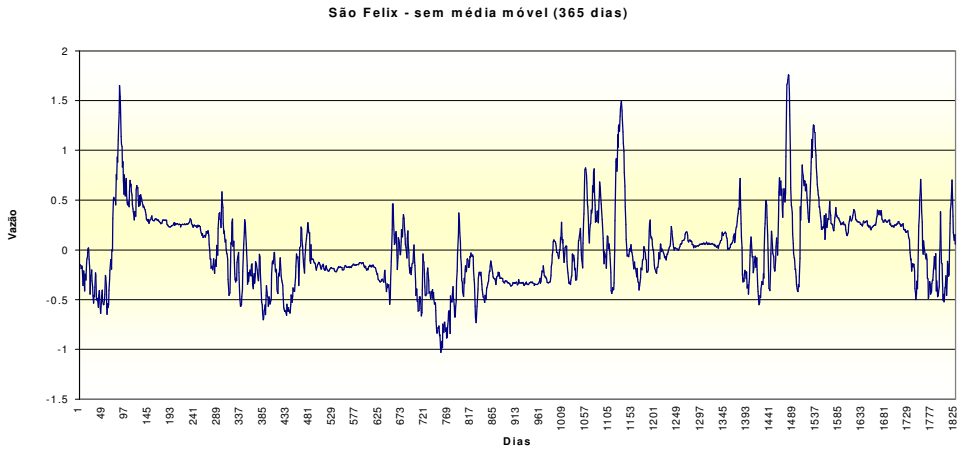


Fig. 30 – Série sem sazonalidade anual.

Retirada de outras componentes senoidais

Após a retirada da sazonalidade anual foi feita na série da Fig. 30 novamente a análise de Fourier, apresentada na Fig. 31. Nota-se a presença ainda de duas componentes dominantes, com valores de período bastante grandes, de 456 e 1826 dias. Estas componentes foram então retiradas, resultando na série de resíduos final, Fig. 32. A análise de Fourier desta série não mostrou novas componentes senoidais à serem retiradas.

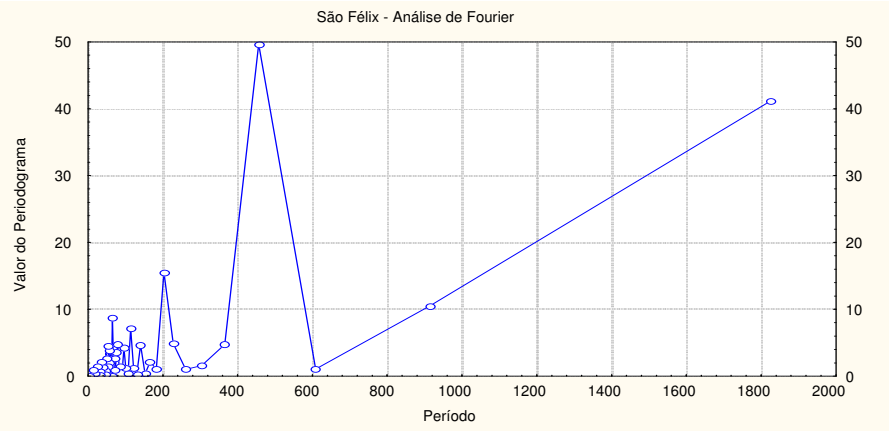


Fig. 31 – Espectrograma da série sem sazonalidade anual.

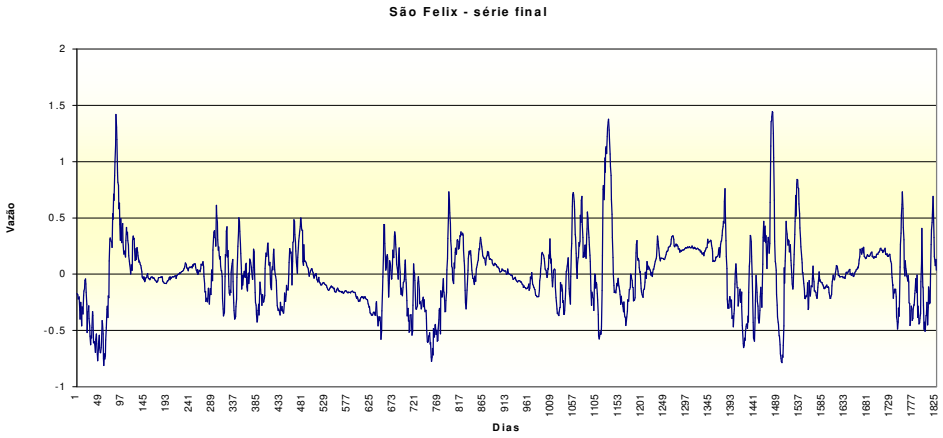


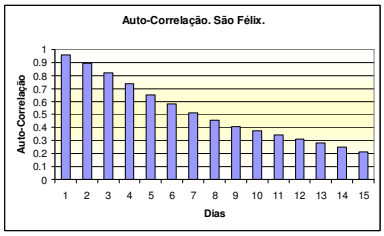
Fig. 32 – Série final de resíduos para a vazão em São Félix.

Pré-processamento das Séries Auxiliares:

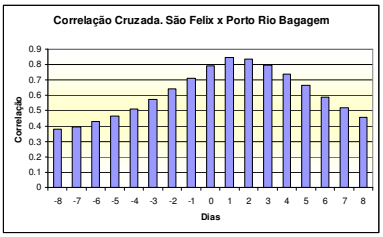
O pré-processamento de todas as séries de vazões foi realizado individualmente para cada série de forma análoga à da série alvo. Os dados da chuva foram apenas normalizados.

Correlações

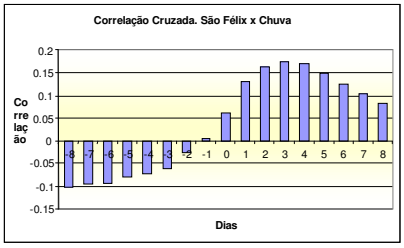
As Fig. 33 abaixo mostram a autocorrelação da vazão em São Félix, e a correlação da vazão em São Félix com a vazão em Rio Bagagem e com a chuva.



(a)



(b)



(c)

Fig. 33 – Correlação das séries residuais de vazões. (a) autocorrelação de São Félix; (b) São Félix com Rio Bagagem e (c) São Félix com a chuva.

A correlação com os demais afluentes tem forma similar àquela com Rio Bagagem. Todas as séries são séries de resíduos, mas mesmo assim as correlações permanecem significativas mesmo para longos atrasos. Como era de se esperar, a chuva demora mais a fazer efeito sobre a vazão dos rios que as vazões a montante. Embora a correlação da chuva seja mais fraca, a chuva aparenta ser importante para sinalizar para a rede se o período é de cheia ou estiagem.

Estrutura, entradas e treinamento da Rede Neural

A partir dos valores das correlações e após diversos testes e optamos por utilizar como entradas da rede neural os valores atrasados de 1 a 5 dias de cada uma das nove séries disponíveis, totalizando 45 entradas. Utilizamos passo de treinamento $\alpha = .0002$ e momento $\beta = .9$. O critério de parada foi o mínimo erro para o conjunto de validação, o que tomava de 1.000 a 15.000 épocas para ser atingido, dependendo do caso. Quando testamos o número de neurônios a utilizar na camada intermediária verificamos com surpresa que com apenas 1 neurônio obtínhamos uma resposta muito boa, logo concluímos que o problema da vazão é bem aproximado por um mapeamento linear. Os resultados obtidos com as condições cima são apresentados a seguir. Para fins de comparação é

apresentado também para cada caso o erro obtido com a decomposição clássica, isto é, modelagem por tendência e sazonalidade apenas.

Resultados:

Previsão para um dia no futuro:

==========

Erro Percentual Médio:

Modelo com Rede Neural, anual: 4.4 %

na cheia: 6.7 %

na estiagem: 2.4 %

Modelo sem Rede Neural, anual: 21.0 %

==========

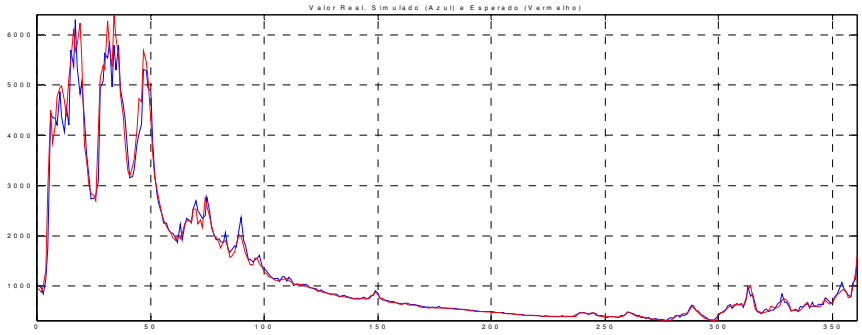


Fig. 34 – Valor real e previsão para 1 dia no futuro. Ano de teste

Previsão para 2 dias no futuro:

==========

Erro Percentual Médio:

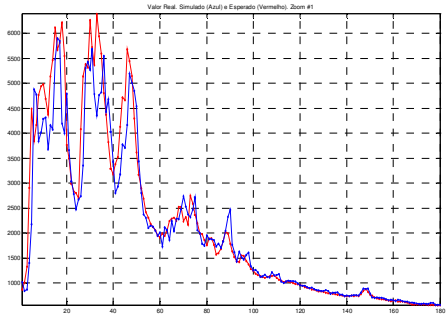
Modelo com Rede Neural, anual: 7.3 %

na cheia: 11.8 %

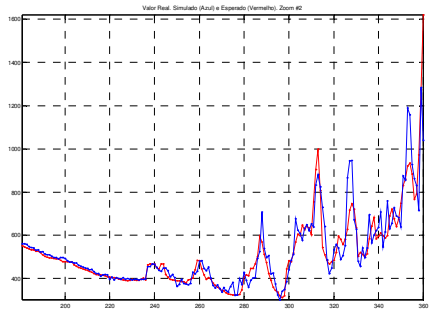
na estiagem: 3.7 %

Modelo sem Rede Neural, anual: 21.0 %

==========



(a)



(b)

Fig. 35 – Valor real e previsão para 2 dias no futuro.

(a) primeiro semestre e (b) segundo semestre do ano de teste

Previsão para 5 dias no futuro:

==========

Erro Percentual Médio:

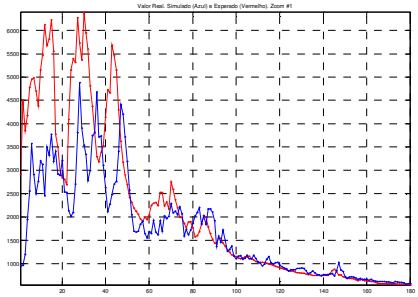
Modelo com Rede Neural, anual: 16.2 %

na cheia: 25.5 %

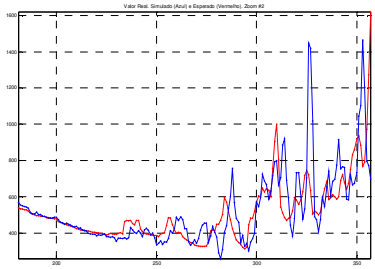
na estiagem: 8.8 %

Modelo sem Rede Neural, anual: 21.0 %

==========



(a)



(b)

Fig. SF11 – Valor real e previsão para 5 dias no futuro.

(a) primeiro semestre e (b) segundo semestre do ano de teste

A partir de cinco dias de antecedência a introdução da rede neural após a decomposição clássica pouco melhora o erro do modelo. O uso da rede, a partir deste momento, não produz melhoria muito significativa na previsão.

A Tabela a seguir apresenta o resumo dos resultados. A coluna Erro *Batch* % apresenta os erros médios para diversos treinamentos da rede neural, e a coluna Erro ótimo % o resultado da melhor rede treinada.

Tabela: Resumo dos Resultados

Dias Futuros	Erro "batch" (%)	Erro ótimo (%)
1	4,7	4,4
2	8,2	7,3
3	10,5	10,2
4	13,5	13,5
5	17,1	16,2

IV - Conclusões:

Redes neurais podem ser uma importante ferramenta auxiliar na modelagem de sistemas dinâmicos e séries temporais.

No caso de sistemas dinâmicos, o modelo discutido neste trabalho é o NARMA, contendo uma rede neural. Operando em modo série-paralelo o NARMA modela fácil e precisamente sistemas dinâmicos não lineares independente do tempo de simulação, mas necessita das saídas passadas da planta para operar. Operando em modo paralelo o modelo se torna autônomo, independente da planta, mas costuma divergir da mesma depois de um número de passos de operação. Este número de passos depende de quão precisa foi a modelagem. A operação em paralelo necessita que, após um certo número de passos, o estado da rede seja atualizado com os valores reais obtidos da planta para que a simulação possa prosseguir com precisão.

Séries Temporais são melhor modeladas iniciando com a decomposição clássica, retirando tendência, sazonalidade, ciclos senoidais, etc., e se necessário transformando matematicamente a série até que a mesma possa ser considerada praticamente estacionária no tempo. Esta série residual é que será modelada, e pode ser vista como a saída de um sistema dinâmico não dissipativo, representável por um modelo NARMA operando em modo série-paralelo e implementado utilizando uma rede neural. As informações atrasadas da série à serem apresentadas á rede são escolhidas por análise de correlação. Informações adicionais tais como dados de séries auxiliares, tempo, amplitude prevista dos sinais da série original, etc., podem também ser fornecidas à rede neural para um melhor mapeamento da série residual. Séries auxiliares, se utilizadas, devem sofrer o mesmo pré-processamento da série alvo antes de serem apresentadas à entrada da rede neural.

Referências Bibliográficas:

Haykin, S., “Neural Networks, A Comprehensive Foundation”, Ch. 13 - 15, Prentice Hall, 1999.

Nascimento, C. L.; Yoneyama, T., “Inteligência Artificial em Contrôle e Automação”, Cap. 13, Ed. Edgard Blucher, São Paulo, 2000.

Luo, F. L.; Unbehauen, R., “Applied Neural Networks for Signal Processing”, Ch. 8, Cambridge Univ. Press, 1998.

“Special Issue on Dynamic Recurrent Neural Systems”, IEEE Trans. on Neural Networks, 5, no. 2, March 94.

Narendra, K. S.; Parthasarathy, K., “Identification and Control of Dynamic Systems Using Neural Networks”, IEEE Trans. on Neural Networks, Vol. 1, no. 1, pg. 4-27, 1990.

Silva, M.F.T., “Identificação de um Sistema Dinâmico através de Redes Neurais Artificiais”, Trabalho de fim de disciplina, EP-UFRJ, 2001. Orientador: L.P. Calôba

Monteiro, J.B., “Identificação de Sistemas Dinâmicos Não Lineares Usando Redes Neurais”, Tese M.Sc., COPPE-UFRJ, 1997. Orientador: L.P. Calôba.

Monteiro, J.B.; Calôba, L.P., “Redes Neurais: Identificação de Sistemas Dinâmicos Não-Lineares com Interferência Reduzida na Operação”, XII Congresso Brasileiro de Automática, Uberlândia, Set. 1998, 6 pg. Também apresentado em:

Monteiro, J.B.; Calôba, L.P., “Nonlinear Dynamic Systems Modeling using Neural Networks with Reduced Interference on the Plant Operation”, Proc. ICSC/IFAC Symposium on Neural Computation, NC’98, Vienna, 1998, 7 pg.;

Monteiro, J.B., Calôba, L.P., “Simulated Annealing: Fast Convergence with Initial Conditions Independence”, Proc. ICSC/IFAC Symposium on Neural Computation NC’98, Vienna, 1998, 7 pg.

Chatfield, C., “The Analysis of Time Series, An Introduction”, Chapman and Hall, 1975.

Morettin, P. A.; Toloi, C. M. C., “Previsão de Séries Temporais”, Atual Editora, São Paulo, 1987.

Ladeira, E.P; França, J.P.M., “Redes Neurais para Cálculo de Vazões”, Trabalho de fim de disciplina, EP-UFRJ, 2000. Orientador: L.P. Calôba

Calôba, G.M., “Cooperação entre Redes Neurais e Técnicas “Clássicas” para Previsão de Demandas”, Trabalho de Fim de Curso de Engenharia, EP-UFRJ, 2001. Orientadores: E. Saliby, L.P.Calôba.

Calôba, G.M., Saliby, E., Calôba, L.P. ““Cooperação entre Redes Neurais e Técnicas “Clássicas” para Previsão de uma Série de Demanda de Cerveja”, Anais do XXXIII Simpósio Brasileiro de Pesquisa Operacional, Campos de Jordão, 2001, 8pg.