3
Ponderação Neural de Experts (NEW)

3.1.
Fundamentos

Ponderação Neural de Experts (NEW - *Neural Expert Weighting*) é um sistema para geração de modelos de ponderação para combinação linear de previsores, baseado em **redes neurais** (seção 3.8). A ideia básica é gerar modelos de ponderação que **relaxem** a dependência dos desempenhos **históricos** dos previsores sendo combinados e agreguem valor aos métodos tradicionais, principalmente na previsões múltiplos passos a frente – agregar valor significa **melhorar** o desempenho ou **encapsular** a complexidade do método original**.** As características teóricas das redes neurais – aproximação universal de funções e resistência ao ruído (robustez) – se mostram adequadas à criação de modelos com este propósito.

Como discutido na seção 1.4, a dependência **exclusiva** dos desempenhos históricos pode ser uma característica limitante para os métodos tradicionais de geração de pesos – nestes métodos, os pesos, múltiplos passos a frente, devem ser estimados exclusivamente com informações de dentro da amostra, já que, fora dela, não há um modelo auxiliar que permita a **atualização** de informações. É principalmente neste ponto que os modelos NEW diferem dos métodos tradicionais, pois geram pesos múltiplos passos a frente que não dependem **exclusivamente** dos desempenhos **exatos** medidos dentro da amostra, mas também de um conjunto de desempenhos **aproximados**, que podem ser calculados ao longo de todo o horizonte de previsão. Para que isto seja possível, introduz-se o conceito de **série de referência**.

Ao considerar-se a série de referência ao invés da série original, troca-se o desempenho exato pelo desempenho aproximado; em outras palavras, trocam-se os erros tomados em relação à série original pelos erros tomados em relação à série de referência, que nada mais é do que uma série gerada por um modelo auxiliar de previsão (seção 3.3). O que se espera com isso é que as redes neurais

do NEW possam aprender funções de ponderação que mapeiem desempenhos aproximados em vetores de pesos. A capacidade de aproximação universal das redes neurais – que são modelos bastante flexíveis de regressão não linear – e sua resistência ao ruído garantem, do ponto de vista teórico, as condições necessárias para esta tarefa. A Figura 8 ilustra a construção de modelos no sistema NEW.

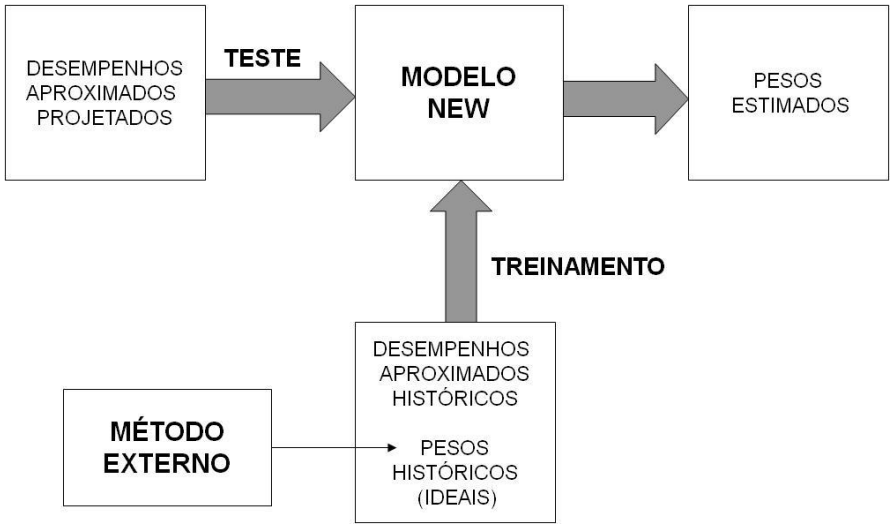

Figura 8 – Diagrama de blocos para um modelo genérico no sistema NEW. As fases de treinamento e teste são mostradas simultaneamente.

Os modelos NEW são sempre construídos em duas fases: **treinamento** e **teste**. Na fase de treinamento, que ocorre antes e uma única vez, o modelo especificado é ajustado com valores históricos de desempenhos aproximados e **pesos ideais**. Na fase de teste, os desempenhos aproximados projetados são mapeados em pesos estimados; se o modelo estiver bem treinado, estes pesos devem ser satisfatórios, i.e., devem agregar valor às combinações lineares múltiplos passos a frente. Nesta linha, a premissa de sucesso de um modelo NEW não está exatamente na qualidade dos previsores ou da série de referência, mas na qualidade do mapeamento da relação referência/previsores (caracterizada pelo cálculo dos desempenhos aproximados) em vetores de pesos. Em última análise, o funcionamento de todo o sistema é baseado nas seguintes variáveis: (i) **previsões** (dentro e fora da amostra), (ii) **série de referência** e (iii) **pesos históricos**.

As previsões dentro e fora da amostra são fornecidas pelos previsores sendo combinados. A cada instante, estas previsões podem ser usadas para calcular desempenhos aproximados, aplicando-se alguma métrica de desempenho (seção 3.2) em relação à série de referência. Dentro da amostra, vetores de desempenhos

aproximados e vetores de pesos históricos (ideais) devem ser **emparceirados** para que se possa realizar o treinamento do modelo NEW em questão. Os pesos históricos devem ser obtidos por algum método externo convenientemente escolhido. Em geral, pode-se adaptar um método tradicional de ponderação para funcionar como método externo, mas nada impede a criação de funções de ponderação empíricas, baseadas na aglutinação de vários métodos (seção 3.4).

Na fase de teste, o modelo treinado deve ser capaz de gerar vetores de pesos múltiplos passos a frente, de maneira dinâmica. Isto é feito mediante a apresentação dos desempenhos aproximados calculados fora da amostra, ou seja, dos desempenhos aproximados projetados. A robustez de generalização do sistema deve ser garantida pelas características teóricas das redes neurais, associadas a procedimentos adequados de pré e pós-processamento das variáveis envolvidas (seções 3.6 e seção 3.7)

3.2. Formulação básica

Nos modelos tradicionais de ponderação o vetor de pesos no instante $\tau + h$ depende, via de regra, dos desempenhos medidos dentro da amostra y^{τ}. Genericamente, esta dependência pode ser representada por uma função envolvendo a realização da série (y) e o vetor de previsões disponíveis ($\hat{\mathbf{y}}$):

$$\hat{\mathbf{w}}_{\tau+h|\tau} = f(y_t, \hat{\mathbf{y}}_{t|t-h}) \qquad (43)$$

onde

$$h \leq H < t \leq \tau$$

Nos modelos NEW a série de interesse $y^{\tau} = y_1, y_2, ..., y_{\tau}$ é substituída pela série de referência $z^{\tau+H} = z_1, z_2, ..., z_{\tau}, z_{\tau+1}, ..., z_{\tau+H}$ gerada por um modelo auxiliar de previsão (seção 3.3). A série z é maior do que a série original y porque contém, além dos valores **ajustados** (instantes 1 a τ), as **previsões** de 1 até H passos a frente. Neste novo paradigma a equação (43) é substituída pela equação (44), com a função G representando uma rede neural devidamente treinada. A formação dos

pares de treinamento dos modelos NEW constitui o núcleo da metodologia, e será aprofundada ao longo desta seção.

$$\hat{\mathbf{w}}^*_{\tau+h|\tau} = G(z_{\tau+h}, \hat{\mathbf{y}}_{\tau+h|\tau}) \tag{44}$$

onde

$h \leq H$

Com a troca da série original por uma de referência, pode-se calcular uma sequência de vetores de desempenhos aproximados (**d**), dentro (45) e fora da amostra (46):

$$\mathbf{d}_{t|t-h} = [d_{t|t-h,1} \quad d_{t|t-h,2} \quad \ldots \quad d_{t|t-h,N}]' \tag{45}$$

onde

$h \leq H < t \leq \tau$

$$\mathbf{d}_{\tau+h|\tau} = [d_{\tau+h|\tau,1} \quad d_{\tau+h|\tau,2} \quad \ldots \quad d_{\tau+h|\tau,N}]' \tag{46}$$

onde

$h \leq H$

Nas equações anteriores, cada componente $d_{T|t,k}$[7] é uma medida de desempenho pontual, calculada para os N previsores disponíveis e podendo ser **absoluta** (47) ou **quadrática** (48):

$$d_{T|t,k} = \left| z_T - \hat{y}_{T|t,k} \right| \tag{47}$$

$$d_{T|t,k} = (z_T - \hat{y}_{T|t,k})^2 \tag{48}$$

onde

[7] Dentro da amostra, o desempenho pontual $d_{T|t,k}$ deve ser escrito como $d_{t|t-h,k}$, $t \leq \tau$, $h \leq H$. Fora da amostra, utiliza-se a notação $d_{\tau+h|\tau,k}$, $h \leq H$.

$$t < T$$
$$k = 1,2,...,N$$

A construção dos pares de treinamento no sistema NEW leva em conta duas grandezas vetoriais, com tantas dimensões quanto forem o número de previsores – (i) média dos desempenhos aproximados e (ii) pesos históricos.

Fixada uma janela de tamanho *v*, pode-se calcular uma sequência de vetores de média dos desempenhos aproximados (**Δ**), dentro (49) e fora da amostra (50). Cada componente (δ) destes vetores é a média do desempenho aproximado (*d*) de um previsor, no intervalo definido pela janela de tempo; as médias são tomadas usando desempenhos em instantes subsequentes, mas com a mesma origem de cálculo: *t* - *h*, se dentro da amostra, ou τ, se fora da amostra.

$$\boldsymbol{\Delta}_{t|t-h}(v) = [\delta_{t|t-h,1} \quad \delta_{t|t-h,2} \quad \ldots \quad \delta_{t|t-h,N}]' = \frac{1}{v}\sum_{i=0}^{v-1}\mathbf{d}_{t-i|t-h} \tag{49}$$

$$onde$$
$$v \le h \le H < t \le \tau$$
$$v \in 1,2,3,...$$

$$\boldsymbol{\Delta}_{\tau+h|\tau}(v) = [\delta_{\tau+h|\tau,1} \quad \delta_{\tau+h|\tau,2} \quad \ldots \quad \delta_{\tau+h|\tau,N}]' = \frac{1}{v}\sum_{i=0}^{v-1}\mathbf{d}_{\tau+h-i|\tau} \tag{50}$$

$$onde$$
$$v \le h \le H$$
$$v \in 1,2,3,...$$

As equações (49) e (50) não precisam necessariamente usar janela de tempo fixa: elas podem ser adaptadas para janela expansiva, de acordo com as equações (51) e (52).

$$\boldsymbol{\Delta}_{t|t-h} = \frac{1}{h}\sum_{i=0}^{h-1}\mathbf{d}_{t-i|t-h} \tag{51}$$

$$onde$$

$$h \leq H < t \leq \tau$$

$$\boldsymbol{\Delta}_{\tau+h|\tau} = \frac{1}{h}\sum_{i=0}^{h-1} \mathbf{d}_{\tau+h-i|\tau} \qquad (52)$$

onde

$$h \leq H$$

Para cada vetor de média $\boldsymbol{\Delta}_{t|t-h}$ dentro da amostra pode-se associar um vetor de pesos históricos $\hat{\mathbf{w}}^*_{t|t-h}$ considerado ideal. A princípio, pesos ideais podem ser derivados de qualquer método racional de ponderação, seja ele tradicional ou empírico (seção 3.4); a única restrição é que o método escolhido seja aplicado considerando-se a série **original** – e não a de **referência** – e as mesmas previsões usadas no cálculo dos desempenhos aproximados $\mathbf{d}_{t|t-h}$ (que por sua vez são usados no cálculo dos desempenhos médios $\boldsymbol{\Delta}_{t|t-h}$). Na equação (53), a função q representa um método genérico para geração de pesos históricos ($\hat{\mathbf{w}}^*_{t|t-h}$).

$$\hat{\mathbf{w}}^*_{t|t-h}(v) = q(y_{t-v+1}, y_{t-v+2}, \ldots, y_t, \hat{\mathbf{y}}_{t-v+1|t-h}, \hat{\mathbf{y}}_{t-v+2|t-h}, \ldots, \hat{\mathbf{y}}_{t|t-h}) \qquad (53)$$

onde

$$v \leq h \leq H < t \leq \tau$$

$$v \in 1,2,3,\ldots$$

Finalmente, a equação (54) traz o formato dos pares de treinamento das redes neurais do sistema NEW, especificados pela associação dos vetores desempenho aproximado médio e peso ideal; quanto maior a relação causal existente entre estas grandezas, maior o potencial de sucesso da metodologia.

$$\left\langle \boldsymbol{\Delta}_{t|t-h}(v), \hat{\mathbf{w}}^*_{t|t-h}(v) \right\rangle \qquad (54)$$

onde

$$v \leq h \leq H < t \leq \tau$$

$$v \in 1,2,3,\ldots$$

Embora a notação seja parecida – $\Delta_{t|t-h}(v)$ e $\Delta_{t|t-h}$ – deve-se observar que, **em geral**, os vetores gerados pela equação (49) são **diferentes** daqueles gerados pela equação (51). Seguem alguns exemplos:

$$\boldsymbol{\Delta}_{4|1}(3) = \frac{1}{3}[\mathbf{d}_{4|1} + \mathbf{d}_{3|1} + \mathbf{d}_{2|1}] = \boldsymbol{\Delta}_{4|1}$$

$$\boldsymbol{\Delta}_{4|1}(1) = \mathbf{d}_{4|1} \neq \boldsymbol{\Delta}_{4|1}$$

$$\boldsymbol{\Delta}_{10|5}(3) = \frac{1}{3}[\mathbf{d}_{10|5} + \mathbf{d}_{9|5} + \mathbf{d}_{8|5}] \neq \boldsymbol{\Delta}_{10|5}$$

$$\boldsymbol{\Delta}_{10|5} = \frac{1}{5}[\mathbf{d}_{10|5} + \mathbf{d}_{9|5} + \mathbf{d}_{8|5} + \mathbf{d}_{7|5} + \mathbf{d}_{6|5}]$$

Para um conjunto de treinamento com τ observações, tomando-se como exemplo a equação (49) e fixando-se uma janela de tamanho v, a aplicação de todos os valores possíveis de t e h leva a uma quantidade de pares de treinamento entre τ - H e (τ - H) x H. Para ilustrar esta afirmação, considera-se um exemplo em que τ = 10 e H = 3; com estes valores, observadas as restrições da equação (49), $v \leq h \leq H < t \leq \tau$, o hiperparâmetro v pode ser escolhido entre 1, 2 ou 3. Fixando-se v = 3, pelas mesmas restrições citadas, h fica constante em 3 e t varia entre 4 e 10, gerando 7 (i.e., τ - H) vetores de desempenho médio: $\Delta_{4|1}(3)$, $\Delta_{5|2}(3)$, $\Delta_{6|3}(3)$, $\Delta_{7|4}(3)$, $\Delta_{8|5}(3)$, $\Delta_{9|6}(3)$ e $\Delta_{10|7}(3)$. Estes 7 valores de $\Delta_{t|t-h}(v)$ definem exatamente 7 pares de treinamento, de acordo com a equação (54). Por outro lado, se for escolhido v = 1, h varia entre 1 e 3, t varia entre 4 e 10 e têm-se 21 (i.e., (τ - H) x H) pares de treinamento, definidos pelos seguintes vetores de desempenho médio: $\Delta_{4|3}(1)$, $\Delta_{4|2}(1)$, $\Delta_{4|1}(1)$, $\Delta_{5|4}(1)$, $\Delta_{5|3}(1)$, $\Delta_{5|2}(1)$, $\Delta_{6|5}(1)$, $\Delta_{6|4}(1)$, $\Delta_{6|3}(1)$, $\Delta_{7|6}(1)$, $\Delta_{7|5}(1)$, $\Delta_{7|4}(1)$, $\Delta_{8|7}(1)$, $\Delta_{8|6}(1)$, $\Delta_{8|5}(1)$, $\Delta_{9|8}(1)$, $\Delta_{9|7}(1)$, $\Delta_{9|6}(1)$, $\Delta_{10|9}(1)$, $\Delta_{10|8}(1)$ e $\Delta_{10|7}(1)$. Com v = 2 têm-se 14 pares de treinamento. Usando a mesma lógica de aplicação da equação (49), é possível concluir que se a equação (51), que contempla janela expansiva, fosse usada em seu lugar, ter-se-ia um número fixo de (τ - H) x H pares de treinamento.

Como deve ser em qualquer problema de regressão, o conjunto de treinamento deve ser projetado de maneira coerente com o objetivo principal – ter um modelo que generalize bem fora da amostra. Fora da amostra, o que se espera

é fornecer um vetor $\boldsymbol{\Delta}_{\tau+h|\tau}$ e obter como resposta um vetor coerente de pesos estimados $\hat{\mathbf{w}}^*_{\tau+h|\tau}$. Matematicamente, tem-se o par de teste da equação (55).

$$\left\langle \boldsymbol{\Delta}_{\tau+h|\tau}(v), \hat{\mathbf{w}}^*_{\tau+h|\tau}(v) \right\rangle \tag{55}$$

onde

$$v \leq h \leq H$$

$$v \in 1,2,3,...$$

Do ponto de vista da implementação, quando um vetor $\boldsymbol{\Delta}_{\tau+h|\tau}$ não puder ser calculado, o modelo de regressão deve retornar um vetor $\hat{\mathbf{w}}^*_{\tau+h|\tau}$ equivalente à média simples (componentes iguais e somando 1).

A Figura 9 traz um diagrama complementar ao exibido na Figura 8, detalhando o funcionamento de um modelo genérico no sistema NEW. As fases de treinamento e teste são exibidas simultaneamente, mas são independentes e ocorrem em momentos distintos. É importante observar o seguinte: desempenhos aproximados (históricos ou projetados) são sempre calculados em relação à série de referência (*z*); pesos (históricos ou projetados) são sempre calculados em relação à série original (*y*).

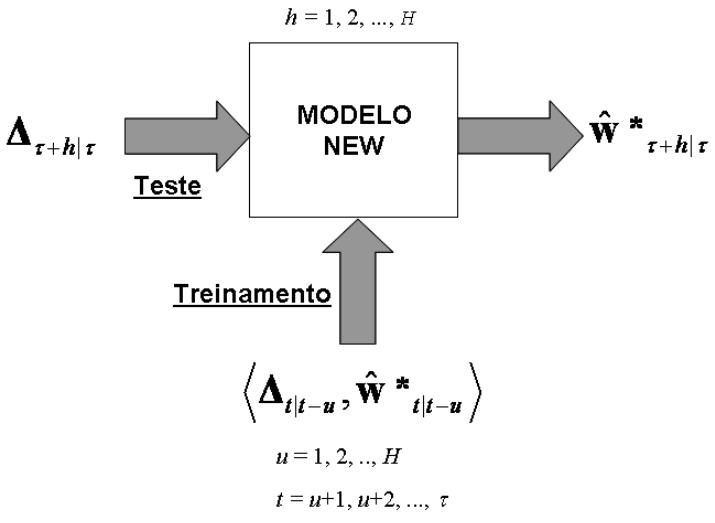

Figura 9 – Diagrama detalhado de um modelo genérico no sistema NEW. Desempenhos são calculados em relação à série de referência. Pesos são calculados em relação à série original.

As próximas seções deste capítulo fornecem detalhes práticos do sistema NEW, contando, quando necessário, com exemplos numéricos.

3.3. Séries de referência

Uma série de referência é tão somente o resultado de um modelo auxiliar de previsão. Em tese, os modelos auxiliares para previsões de referência podem ser quaisquer; contudo, em termos práticos, eles devem ser obtidos por metodologias simples, que não demandem ajustes subjetivos. O motivo para isso é justificável pelo paradigma seleção *versus* combinação, explorado na seção 1.3: investir tempo e conhecimento na busca de um modelo auxiliar ótimo teria custo equivalente à tentativa de seleção do melhor modelo de previsão, e o foco do sistema NEW, por outro lado, é diminuir o risco por combinação ponderada.

Para formação das séries de referência, os quatro modelos listados a seguir foram testados neste trabalho (considerando séries mensais com período = 12). Os modelos 1, 2 e 3 são chamados, respectivamente, de SALY1, SALY2 e SALY3 – SALY é acrônimo de *Same As Last Year*. O modelo 4 foi abreviado ao longo do texto pela sigla DEC.

1. Repetição do último período:
    a. $z_t = y_{t-12}$, se $12 < t \leq \tau + 12$;
    b. $z_t = z_{t-12}$, se $t > \tau + 12$.
2. Repetição do último período com crescimento sazonal (γ) mais recente:
    a. $z_t = y_{t-12} . \gamma_{t|t-12}$, se $12 < t \leq \tau + 12$;
    b. $z_t = z_{t-12} . \gamma_{t|t-12}$, se $t > \tau + 12$.
3. Repetição do último período com crescimento sazonal médio (γ^*):
    a. $z_t = y_{t-12} . \gamma^*_{t|t-12}$, se $12 < t \leq \tau + 12$;
    b. $z_t = z_{t-12} . \gamma^*_{t|t-12}$, se $t > \tau + 12$.
4. Decomposição clássica com nível, tendência linear e sazonalidade multiplicativa (Apêndice A).

Dado o problema em questão, a determinação da série de referência mais adequada deve ser feita de maneira criteriosa. Por exemplo, (i) pode-se testar

modelos NEW baseados em diversas referências, elegendo o par modelo/ referência que tenha fornecido o menor erro de previsão combinada, medido em um conjunto de validação; outra possibilidade, mais direta, é (ii) escolher, dentro da amostra, a referência que tenha a menor **correlação conjunta** com os previsores sendo combinados, entendendo-se por correlação conjunta um valor único para medir a relação linear entre uma dada referência e o conjunto de todos os previsores[8]. Este último critério tem sentido, pois como mencionado anteriormente (seção 3.1), a premissa de sucesso de um modelo NEW está na qualidade do mapeamento da relação referência/previsores em vetores de pesos; admitindo que este mapeamento, em teoria, possa sempre ser aprendido por uma rede neural bem dimensionada, quanto menor a correlação conjunta observada maior a diversidade dos componentes e, portanto, menor o risco global do sistema (seções 2.1 e 2.2.2). Finalmente, uma terceira possibilidade para escolha da série de referência, mais completa e utilizada nos experimentos do capítulo 4, é (iii) usar um **critério misto**, combinando (i) e (ii).

Embora fora do escopo final deste trabalho, uma alternativa para geração em massa de séries de referência seria por simulação de Monte Carlo (Bishop, 1995; Ross, 2006). Por exemplo, pode-se selecionar uma série de referência básica e replicá-la *R* vezes, sendo cada replicação equivalente à série original multiplicada por um fator de erro, sorteado de uma distribuição normal. A Figura 10 ilustra este procedimento.

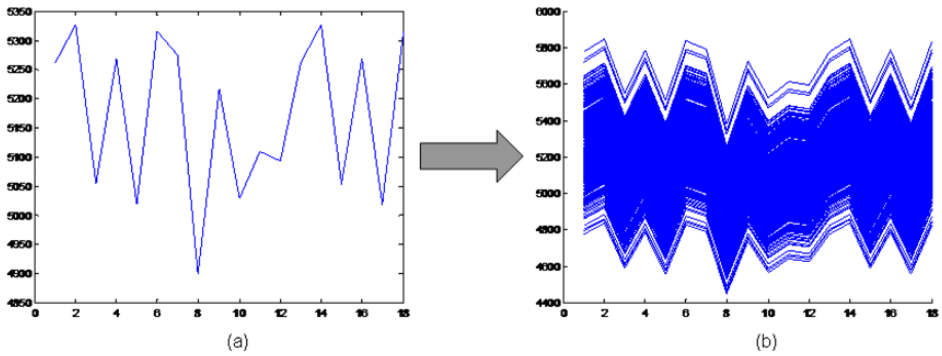

Figura 10 – (a) Série de referência original; (b)1000 replicações Monte Carlo.

[8] Neste trabalho, a correlação conjunta é aproximada pela média das *N* correlações de erro referência/previsor observadas (seção 4.4.4.1).

Por fim, vale salientar que se a série de referência (*z*) usada no treinamento de um modelo NEW for igual à série real (*y*), os desempenhos calculados passam a ser exatos (e não aproximados) e a relação causal com os pesos históricos aumenta; contudo, este procedimento só tem sentido a título de *benchmarking*, já que por definição não se conhece a realização da série fora da amostra, inviabilizando testes coerentes. Em outro caso extremo, poder-se-ia considerar uma série de referência nula. Isso significaria dizer que os modelos gerados pelo NEW estariam sendo treinados para gerar pesos considerando apenas os valores das previsões sendo combinadas, uma tarefa sem fundamentação teórica.

3.4.Pesos históricos

A abordagem mais direta para obtenção de pesos históricos é adaptar um dos métodos citados na seção 2.4. As equações a seguir mostram como os métodos de mínimos quadrados restritos – (56) e (57) – Bates & Granger simples (BG1) (58) e AFTER – (59) a (61) – devem ser adaptados para funcionar como geradores de pesos de treinamento para o sistema NEW. Nestas adaptações, são duas as mudanças a serem observadas: (i) para cada vetor $\hat{\mathbf{w}}^*_{t|t-h}$ a **origem** das previsões usadas no cálculo é sempre a mesma – *t* - *h* – e (ii) as equações geradoras de pesos incluem **defasagem** (*lag*) **zero**, ou seja, o cálculo de um vetor de pesos para o instante *t* leva em conta valores da série original **menores ou iguais** a *t* (e não apenas **menores** que *t*). Assim como na seção 2.4, as equações (56) a (61) podem ser adaptadas para janela expansiva (neste caso, eliminando o "(*v*)" no lado esquerdo e substituindo *v* por *h* no lado direito).

$$\hat{\mathbf{w}}^*_{t|t-h}(v) = \begin{bmatrix} \hat{w}^*_{t|t-h,1} \\ \hat{w}^*_{t|t-h,2} \\ \vdots \\ \hat{w}^*_{t|t-h,N} \end{bmatrix} = \min_{w^*} \sum_{i=0}^{v-1} \left(y_{t-i} - \sum_{k=1}^{N} w^*_{t-i|t-h,k} \cdot \hat{y}_{t-i|t-h,k} \right)^2 \qquad (56)$$

sujeito a

$$\sum_{k=1}^{N} w^*_{t|t-h,k} = 1 \quad e \quad w^*_{t|t-h,k} \geq 0 \qquad (57)$$

$$v \leq h \leq H < t \leq \tau$$

$$v \in 1,2,3,...$$

$$\hat{w}^*_{t|t-h,k}(v) = \frac{\left[v^{-1} \sum_{i=0}^{v-1} (y_{t-i} - \hat{y}_{t-i|t-h,k})^2 \right]^{-1}}{\sum_{j=1}^{N} \left[v^{-1} \sum_{i=0}^{v-1} (y_{t-i} - \hat{y}_{t-i|t-h,j})^2 \right]^{-1}} \tag{58}$$

onde

$$v \leq h \leq H < t \leq \tau$$

$$v \in 1,2,3,...$$

$$\hat{w}^*_{t|t-h,k}(\hat{\sigma}^*) = \frac{\hat{w}_{t-1|t-h,k} \hat{\sigma}^{*-1/2}_{t|t-h,k} \exp\{-[(y_t - \hat{y}_{t|t-h,k})^2 / 2\hat{\sigma}^*_{t|t-h,k}]\}}{\sum_{j=1}^{N} \hat{w}_{t-1|t-h,j} \hat{\sigma}^{*-1/2}_{t|t-h,j} \exp\{-[(y_t - \hat{y}_{t|t-h,j})^2 / 2\hat{\sigma}^*_{t|t-h,j}]\}} \tag{59}$$

onde

$$\hat{\sigma}^*_{t|t-h,k} = \hat{\sigma}^*_{t|t-h,k}(v) = v^{-1} \sum_{i=0}^{v-1} \varepsilon^2_{t-i|t-h,k} \tag{60}$$

$$\varepsilon_{t|t-h,k} = (y_t - \hat{y}_{t|t-h,k}) \tag{61}$$

$$h \leq H < t \leq \tau$$

A forma alternativa para geração de pesos de treinamento é empregar uma abordagem híbrida, aglutinando um conjunto de métodos geradores (ao invés de usar um único) ou incorporando alterações manuais nas funções de ponderação. Dentro da primeira abordagem - usar mais de um método de formação - uma estratégia possível é a seguinte:

1. No conjunto de treinamento, formar τ - H blocos reunindo previsões de 1 até H passos frente, sendo cada bloco definido pela origem t das previsões ($t \leq \tau$ - H);
2. Para cada bloco, associar o trecho correspondente da série real (y_t, y_{t+1}, ..., y_{t+H}) e aplicar todos os métodos de ponderação candidatos individualmente;
3. Associar, a cada bloco, o método candidato que leve ao menor erro de **previsão combinada** (realização da série **menos** previsão do modelo de combinação).

Assim, ao se formar os pares de treinamento de um modelo NEW, o método de ponderação usado variará com o **bloco** de previsões. Esta estratégia é chamada neste trabalho de *pool* de métodos.

3.5. Conjuntos de treinamento

Para exemplificar a formação de conjuntos de treinamento no sistema NEW considera-se uma série original *y* e uma série de referência *z*, exibidas na Figura 11 e na Tabela 1. Neste exemplo, tem-se um conjunto de treinamento de tamanho $\tau = 24$. Na prática a série *z* teria *H* observações a mais do que a série *y* (*H* é o horizonte máximo de previsão), mas do ponto vista da formação dos conjuntos de treinamento, estes dados não têm relevância.

Admitindo-se $H = 12$, pode-se formar, para cada previsor admitido na combinação, até 12 blocos com previsões de 1 até 12 passos a frente. Em outras palavras, na amostra de tamanho $\tau = 24$ pode-se formar até $\tau - H$ blocos de previsões $\hat{y}_{t|t-h,k}$, k variando de 1 até N. A Tabela 2 considera um modelo de combinação com $N = 2$ e exibe os dois primeiros blocos de previsões.

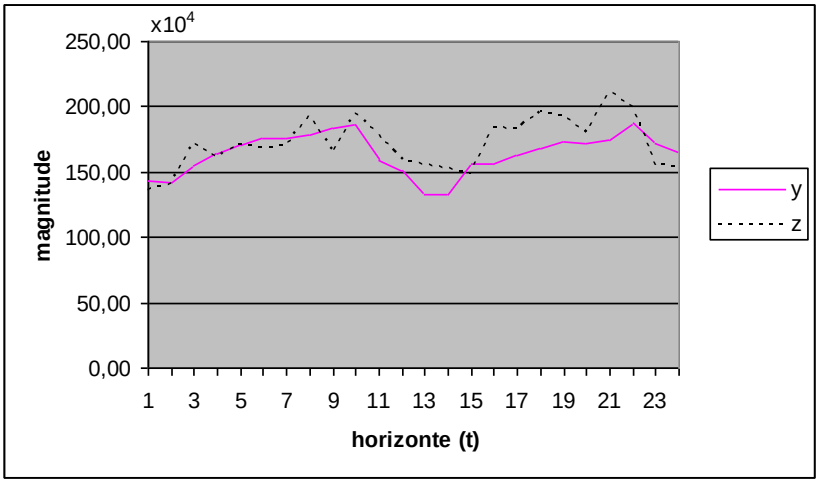

Figura 11 – Série original (*y*) e série de referência (*z*). Na figura, a série *z* está limitada ao intervalo dentro da amostra, mas, na prática, ela se entende por todo o horizonte de previsão.

Tabela 1 – Valores traçados na Figura 11

t	y (x10^4)	z (x10^4)
1	143.14	134.82
2	141.28	141.04
3	154.36	171.91
4	164.28	160.23
5	171.09	170.13
6	176.11	168.37
7	175.39	168.66
8	178.15	192.53
9	183.56	163.84
10	186.47	193.71
11	159.31	178.53
12	150.91	159.26
13	133.06	154.89
14	132.24	152.35
15	156.30	148.55
16	156.86	183.01
17	162.85	181.77
18	167.88	195.20
19	173.81	193.22
20	171.30	179.77
21	174.43	210.29
22	187.10	199.05
23	171.93	154.99
24	165.64	153.40

Tabela 2 – Blocos de previsões até 12 passos a frente para dois previsores, considerando como origem os instantes 1 e 2 (total de 24 previsões)

t	t-h	$\hat{y}_{t\|t-h,1}$	$\hat{y}_{t\|t-h,2}$	y (x10^4)	z (x10^4)
2	1	149.79	123.75	141.28	141.04
3	1	174.13	148.09	154.36	171.91
4	1	166.81	140.77	164.28	160.23
5	1	176.22	150.18	171.09	170.13
6	1	176.55	150.51	176.11	168.37
7	1	179.81	153.77	175.39	168.66
8	1	191.71	165.67	178.15	192.53
9	1	183.26	157.22	183.56	163.84
10	1	190.77	164.72	186.47	193.71
11	1	178.69	152.65	159.31	178.53
12	1	167.68	141.64	150.91	159.26
3	2	157.98	131.94	154.36	171.91
4	2	149.79	123.75	164.28	160.23
5	2	174.13	148.09	171.09	170.13
6	2	166.81	140.77	176.11	168.37
7	2	176.22	150.18	175.39	168.66
8	2	176.55	150.51	178.15	192.53
9	2	179.81	153.77	183.56	163.84
10	2	191.71	165.67	186.47	193.71
11	2	183.26	157.22	159.31	178.53
12	2	190.77	164.72	150.91	159.26
13	2	178.69	152.65	133.06	154.89

Os pares de treinamento para o modelo NEW pretendido devem ser obtidos com os dados da Tabela 2, de acordo com a equação (54). Neste ponto, o projetista deve estabelecer os seguintes hiperparâmetros:

1. Tipo de desempenho aproximado: absoluto (47) ou quadrático (48). Ao longo do texto, os tipos de desempenho são referenciados, respectivamente, pelas siglas MAD e MSE.
2. Método de geração de pesos históricos (seção3.4);
3. Tamanho da janela de tempo.

Considerando-se como método de geração de pesos históricos o BG1 (58), as Tabelas 3 a 6 exibem, para diferentes tipos de desempenho e janelas de tempo, os pares de treinamento formados com os dois blocos de previsões da Tabela 2. Linhas vazias nas tabelas ocorrem nos casos em que se tem janela de tempo fixa (de tamanho v): nestes casos, como se observa nas equações (49) e (58), as grandezas δ e w^* não são calculáveis para instantes $t < v$.

Tabela 3 – Pares de treinamento: desempenho MAD, janela expansiva

#par	$\delta_{t\|t-h,1}$ (x10^4)	$\delta_{t\|t-h,2}$ (x10^4)	$\hat{w}^*_{t\|t-h,1}$	$\hat{w}^*_{t\|t-h,2}$
1	8.75	17.29	0.81	0.19
2	5.49	20.55	0.43	0.57
3	5.85	20.19	0.66	0.34
4	5.91	20.13	0.73	0.27
5	6.37	19.68	0.80	0.20
6	7.16	18.88	0.83	0.17
7	6.26	20.02	0.79	0.21
8	7.90	18.34	0.83	0.17
9	7.35	19.53	0.84	0.16
10	6.63	20.16	0.78	0.22
11	6.80	19.93	0.74	0.26
12	6.49	20.18	0.66	0.34
13	3.18	22.86	0.06	0.94
14	5.31	20.73	0.55	0.45
15	5.86	20.18	0.66	0.34
16	6.65	19.39	0.76	0.24
17	7.73	18.31	0.80	0.20
18	6.46	19.58	0.75	0.25
19	8.44	17.60	0.79	0.21
20	7.63	18.89	0.81	0.19
21	6.91	19.55	0.74	0.26
22	7.17	19.26	0.69	0.31
23	6.89	19.50	0.60	0.40
24	6.63	19.73	0.55	0.45

Teoricamente, os conjuntos de treinamento das Tabelas 3 a 6 poderiam ser usados, como estão, para ajustar as redes neurais do sistema NEW. Contudo, observa-se na prática a necessidade de um pré-processamento específico nestes

conjuntos, sob pena de perda acentuada de desempenho e mesmo não convergência dos algoritmos iterativos responsáveis pelo ajuste. Este assunto é desenvolvido na seção 3.6.

Tabela 4 – Pares de treinamento: desempenho MSE, janela expansiva

#par	$\delta_{t\|t-h,1}$ (x10^8)	$\delta_{t\|t-h,2}$ (x10^8)	$\hat{w}^*_{t\|t-h,1}$	$\hat{w}^*_{t\|t-h,2}$
1	76.57	298.96	0.81	0.19
2	40.76	433.05	0.43	0.57
3	41.61	414.93	0.66	0.34
4	40.47	410.74	0.73	0.27
5	45.76	392.39	0.80	0.20
6	58.84	363.96	0.83	0.17
7	50.53	415.04	0.79	0.21
8	91.34	368.65	0.83	0.17
9	82.16	421.07	0.84	0.16
10	73.95	445.90	0.78	0.22
11	73.67	433.59	0.74	0.26
12	68.33	441.35	0.66	0.34
13	10.09	522.80	0.06	0.94
14	32.72	434.41	0.55	0.45
15	37.96	410.97	0.66	0.34
16	48.88	380.51	0.76	0.24
17	67.99	343.74	0.80	0.20
18	56.66	398.37	0.75	0.25
19	107.75	346.08	0.79	0.21
20	94.75	400.56	0.81	0.19
21	84.38	424.69	0.74	0.26
22	84.87	409.75	0.69	0.31
23	78.72	416.05	0.60	0.40
24	73.31	422.93	0.55	0.45

Tabela 5 – Pares de treinamento: desempenho MAD, janela 2

#par	$\delta_{t\|t-h,1}$ (x10^4)	$\delta_{t\|t-h,2}$ (x10^4)	$\hat{w}^*_{t\|t-h,1}$	$\hat{w}^*_{t\|t-h,2}$
1				
2	5.49	20.55	0.43	0.57
3	4.40	21.64	0.60	0.40
4	6.33	19.71	0.97	0.03
5	7.13	18.91	0.98	0.02
6	9.66	16.38	0.98	0.02
7	5.98	20.88	0.75	0.25
8	10.12	16.74	0.82	0.18
9	11.18	17.81	0.98	0.02
10	1.56	27.43	0.57	0.43
11	4.30	21.75	0.17	0.83
12	5.76	20.28	0.09	0.91
13	5.49	20.55	0.43	0.57
14				
15	5.31	20.73	0.55	0.45
16	7.20	18.84	0.95	0.05
17	8.00	18.04	0.96	0.04
18	10.53	15.51	0.97	0.03
19	6.07	19.97	0.70	0.30
20	10.24	15.80	0.79	0.21
21	11.14	16.82	0.97	0.03
22	1.55	26.41	0.51	0.49
23	5.32	20.72	0.12	0.88
24	6.80	19.24	0.06	0.94

Tabela 6 – Pares de treinamento: desempenho MSE, janela 2

#par	$\delta_{t\|t-h,1}$ (x10^8)	$\delta_{t\|t-h,2}$ (x10^8)	$\hat{w}^*_{t\|t-h,1}$	$\hat{w}^*_{t\|t-h,2}$
1				
2	40.76	433.05	0.43	0.57
3	24.13	472.92	0.60	0.40
4	40.18	388.43	0.97	0.03
5	51.99	358.57	0.98	0.02
6	95.58	270.41	0.98	0.02
7	62.45	471.68	0.75	0.25
8	188.85	382.70	0.82	0.18
9	192.86	442.16	0.98	0.02
10	4.36	754.94	0.57	0.43
11	35.49	489.91	0.17	0.83
12	40.25	418.57	0.09	0.91
13				
14	32.72	434.41	0.55	0.45
15	51.89	355.05	0.95	0.05
16	65.05	326.61	0.96	0.04
17	113.05	242.89	0.97	0.03
18	72.22	434.10	0.70	0.30
19	207.15	351.94	0.79	0.21
20	208.99	407.14	0.97	0.03
21	2.55	699.83	0.51	0.49
22	45.36	446.48	0.12	0.88
23	53.28	377.18	0.06	0.94
24	15.51	488.87	0.00	1.00

Neste trabalho, considera-se como pré-processamento a atividade de normalizar (transformar) as variáveis de entrada[9] em um modelo NEW. Esta atividade tem grande influência prática no que se refere ao uso de redes neurais; dependendo do tipo de normalização realizada (ou da ausência de normalização), o treinamento será mais ou menos estável, e o teste mais ou menos robusto (NEURAL-NETS FAQ, 2011).

[9] Variáveis de entrada são as variáveis no lado esquerdo do par de treinamento <**Δ**;**w***>.

3.6.1.
Normalização padrão

A normalização padrão, também chamada de **padronização**, é comumente usada na prática (Reed & Marks, 1999; Mathworks, 2010b); ela leva em conta a distribuição estatística dos vetores de desempenhos aproximados médios, $\Delta_{t|t-h} = [\delta_{t|t-h,1} \ \ \delta_{t|t-h,2} \ \ ... \ \delta_{t|t-h,N}]'$. Cada componente transformado $\delta'_{t|t-h,k}$ é obtido pela subtração da média ($\bar{\delta}_k$) e posterior divisão pelo respectivo desvio padrão (σ_k):

$$\delta'_{t|t-h,k} = \frac{\delta_{t|t-h,k} - \bar{\delta}_k}{\sigma_k} \tag{62}$$

onde

$$\bar{\delta}_k = E(\delta_{t|t-h,k}) \tag{63}$$

$$\sigma_k = \sqrt{E([\delta_{t|t-h,k} - \bar{\delta}_k]^2)} \tag{64}$$

Os valores $\bar{\delta}_k$ e σ_k devem ser calculados na fase de treinamento e guardados para transformar as entradas na fase de teste. Aplicada no exemplo da Tabela 3, a normalização padrão gera o conjunto de treinamento exibido na Tabela 7.

3.6.2.
Normalização soma-1

Esta normalização alternativa é inspirada nas ideias do procedimento BG1 para geração tradicional de pesos (seção 2.4.4). Ela deve ser aplicada individualmente a cada vetor $\Delta_{t|t-h} = [\delta_{t|t-h,1} \ \ \delta_{t|t-h,2} \ \ ... \ \delta_{t|t-h,N}]'$, simplesmente garantindo que a soma dos seus componentes seja 1; não há qualquer consideração estatística sobre o conjunto de vetores. Matematicamente, a transformação é a seguinte:

$$\delta'_{t|t-h,k} = \frac{\delta_{t|t-h,k}}{\sum_{k=1}^{N} \delta_{t|t-h,k}} \tag{65}$$

A normalização soma-1 estabelece, a cada instante *t*, uma escala relativa entre os desempenhos aproximados dos previsores sendo combinados. Aplicada no exemplo da Tabela 3, ela gera o conjunto de treinamento exibido na Tabela 8.

Tabela 7 – Normalização padrão

#entrada	$\delta_{t\|t-h,1}$ (x10^4)	$\delta_{t\|t-h,2}$ (x10^4)	$\delta'_{t\|t-h,1}$	$\delta'_{t\|t-h,2}$
1	8.75	17.29	1.84	-2.08
2	5.49	20.55	-1.03	0.85
3	5.85	20.19	-0.71	0.53
4	5.91	20.13	-0.66	0.47
5	6.37	19.68	-0.26	0.07
6	7.16	18.88	0.44	-0.65
7	6.26	20.02	-0.35	0.37
8	7.90	18.34	1.09	-1.13
9	7.35	19.53	0.61	-0.07
10	6.63	20.16	-0.02	0.50
11	6.80	19.93	0.12	0.29
12	6.49	20.18	-0.15	0.52
13	3.18	22.86	-3.06	2.93
14	5.31	20.73	-1.19	1.01
15	5.86	20.18	-0.70	0.52
16	6.65	19.39	0.00	-0.19
17	7.73	18.31	0.94	-1.16
18	6.46	19.58	-0.17	-0.02
19	8.44	17.60	1.57	-1.80
20	7.63	18.89	0.85	-0.64
21	6.91	19.55	0.22	-0.04
22	7.17	19.26	0.45	-0.31
23	6.89	19.50	0.21	-0.10
24	6.63	19.73	-0.03	0.12
$\bar{\delta}_1 = 6.66\text{x}10^4,\ \bar{\delta}_2 = 19.60\text{x}10^4$				
$\sigma_1 = 1.14\text{x}10^4,\ \sigma_2 = 1.11\text{x}10^4$				

Tabela 8 – Normalização soma-1

#entrada	$\delta_{t\|t-h,1}$ (x10^4)	$\delta_{t\|t-h,2}$ (x10^4)	$\delta'_{t\|t-h,1}$	$\delta'_{t\|t-h,2}$
1	8.75	17.29	0.34	0.66
2	5.49	20.55	0.21	0.79
3	5.85	20.19	0.22	0.78
4	5.91	20.13	0.23	0.77
5	6.37	19.68	0.24	0.76
6	7.16	18.88	0.28	0.72
7	6.26	20.02	0.24	0.76
8	7.90	18.34	0.30	0.70
9	7.35	19.53	0.27	0.73
10	6.63	20.16	0.25	0.75
11	6.80	19.93	0.25	0.75
12	6.49	20.18	0.24	0.76
13	3.18	22.86	0.12	0.88
14	5.31	20.73	0.20	0.80
15	5.86	20.18	0.22	0.78
16	6.65	19.39	0.26	0.74
17	7.73	18.31	0.30	0.70
18	6.46	19.58	0.25	0.75
19	8.44	17.60	0.32	0.68
20	7.63	18.89	0.29	0.71
21	6.91	19.55	0.26	0.74
22	7.17	19.26	0.27	0.73
23	6.89	19.50	0.26	0.74
24	6.63	19.73	0.25	0.75

Como discutido na seção 2.4, muitos métodos tradicionais para geração de pesos garantem vetores $\hat{\mathbf{w}}_{\tau+h|\tau} = [\hat{w}_{\tau+h|\tau,1} \ \hat{w}_{\tau+h|\tau,2} \ \ldots \ \hat{w}_{\tau+h|\tau,N}]'$ com componentes somando 1; isto porque, no caso linear, a restrição de "somar 1" garante que a previsão consensual seja não tendenciosa, se os previsores combinados assim o forem (seção 2.4.3). Deste modo, sempre que um modelo NEW for treinado com pesos históricos que atendam à restrição de não tendenciosidade, é válido que os vetores $\hat{\mathbf{w}}^*_{\tau+h|\tau}$ por ele gerados na fase de teste sejam **pós-processados**, de maneira que a restrição permaneça atendida. Esta atividade é chamada neste

trabalho de **reconciliação de pesos**, e consiste em transformar os componentes de $\hat{\mathbf{w}}^*_{\tau+h|\tau}$ de acordo com a equação (66).

$$\hat{w}^*_{\tau+h|\tau,k} = \frac{\hat{w}^*_{\tau+h|\tau,k}}{\sum_{k=1}^{N} \hat{w}^*_{\tau+h|\tau,k}} \tag{66}$$

3.8. Redes neurais

Em última análise, o sistema NEW é uma metodologia para construção de modelos de ponderação baseados em redes neurais *multilayer perceptron* (MLP) (Bishop, 1995; Haykin, 1998; Reed & Marks, 1999). Do ponto de vista estatístico, uma rede MLP é um modelo de regressão não linear, com grande potencial para aproximação de funções; em tese, esta característica dá a flexibilidade necessária para criação de funções de ponderação complexas. Do ponto de vista da inteligência computacional, redes MLP são modelos de dados com capacidade de aprendizado e generalização, permitindo a ponderação de previsores baseada em (bons) exemplos históricos.

No modelo NEW mais geral, cada componente do vetor de ponderação $\hat{\mathbf{w}}^*_{\tau+h|\tau} = [\hat{w}^*_{\tau+h|\tau,1} \;\; \hat{w}^*_{\tau+h|\tau,2} \; ... \; \hat{w}^*_{\tau+h|\tau,N}]'$ é definido pela equação (67). Nesta equação, a função g é uma rede neural com saída única; ela representa **parcialmente** a rede neural G – equação (44) – que tem uma saída para cada um dos (N) componentes de $\hat{\mathbf{w}}^*_{\tau+h|\tau}$.

$$\hat{w}^*_{\tau+h|\tau,k} = g(\boldsymbol{\Delta}_{\tau+h|\tau}, \mathbf{b}) = \sum_{i=1}^{p} \beta_i . tanh\left(\sum_{j=1}^{N} \beta_{2p+(i-1)N+j} \delta_{\tau+h|\tau,j} + \beta_{p+i}\right) + \beta_0 \tag{67}$$

onde

$$\boldsymbol{\Delta}_{\tau+h|\tau} = [\delta_{\tau+h|\tau,1} \quad \delta_{\tau+h|\tau,2} \quad ... \quad \delta_{\tau+h|\tau,N}]' \tag{68}$$

$$\mathbf{b} = [\beta_0 \quad \beta_1 \quad ... \quad \beta_{p(N+2)}]' \tag{69}$$

$$tanh(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} \tag{70}$$

Nas equações (67) e (69), o vetor **b** reúne os $p(N+2)+1$ parâmetros – chamados de pesos **sinápticos** – a serem ajustados no treinamento da rede neural (considerada até aqui, por simplificação, com saída única). O limite superior p no primeiro somatório da equação (67), referido na literatura como número de neurônios na camada escondida, é um hiperparâmetro a ser determinado pela política de treinamento escolhida (seção 3.9). Nas equações (67) e (70), a função tangente hiperbólica (*tanh*) tem propriedades que validam o chamado teorema da aproximação universal (Cybenko, 1989), teorema que suporta o uso das redes neurais como modelos robustos de regressão não linear.

A Figura 12 exibe uma representação gráfica para a equação (67); cada conexão (arco) representa um peso sináptico e cada um dos escalares fixos em +1 é chamado de *bias* (viés). A contagem das conexões existentes – nem todas estão desenhadas na figura – dá a quantidade de parâmetros da rede neural com saída única: $(N+1)p + (p+1)1 = p(N+2)+1$.

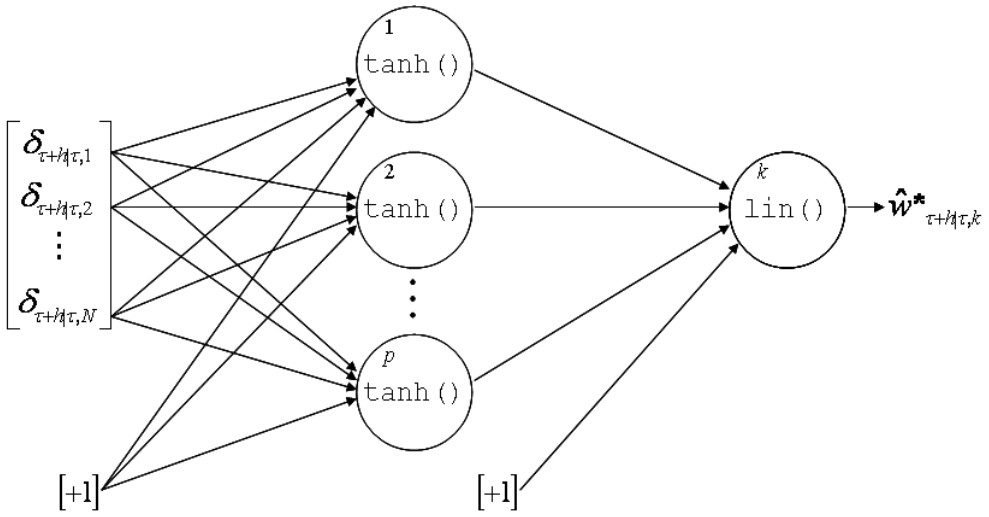

Figura 12 – Representação gráfica da rede neural com saída única. Os círculos de 1 até p são os neurônios da camada escondida; o círculo k é o neurônio de saída, com função de ativação linear. Por simplicidade, nem todas as conexões existentes foram exibidas.

Nos modelos NEW completos, as redes neurais (G) não têm apenas uma saída; elas têm N saídas, uma para cada componente do vetor $\hat{\mathbf{w}}^*_{\tau+h|\tau}$, como exibe a Figura 13. Contando as novas conexões, é possível notar que o número final de parâmetros da rede aumenta, passando de p(N+2)+1 para $2pN + p + N$.

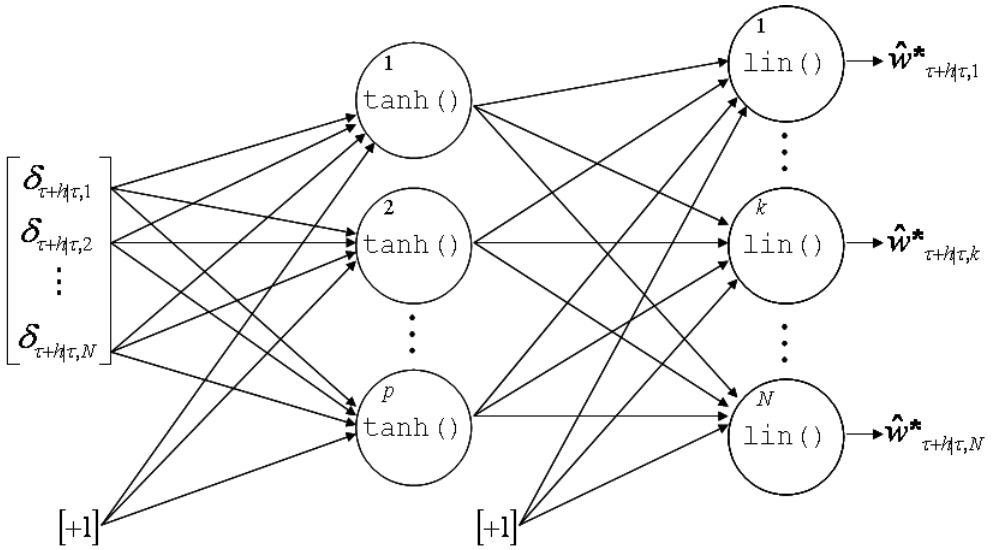

Figura 13 – Representação gráfica da rede neural com múltiplas saídas. Os círculos de 1 até *p* são os neurônios da camada escondida; os círculos de 1 até *N* são os neurônios de saída, com função de ativação linear. Por simplicidade, nem todas as conexões existentes foram exibidas.

Nos modelos apresentados até o momento (Figuras 12 e 13), os valores de $\hat{w}^*_{\tau+h|\tau,k}$ podem variar livremente em $\Re$; em outras palavras, diz-se que a função de transferência nas saídas da rede neural é linear ($lin(u)=u$). Esta característica é desejável quando da estimação de valores de pesos que estejam fora do intervalo [0,1], como por exemplo, aqueles gerados pelos métodos de mínimos quadrados irrestritos ou BG2 (seções 2.4.2 e 2.4.5). Por outro lado, como mencionado na seção 2.4.3, combinações convexas – com pesos no intervalo [0,1] e somando 1 – tendem a ser mais estáveis do que as irrestritas; além disso, é intuitivo notar que o uso de saídas limitadas facilita, do ponto de vista numérico, o processo de aprendizado das redes neurais. Para garantir componentes $\hat{w}^*_{\tau+h|\tau,k}$ no intervalo [0,1], a equação (67) deve ser substituída pela equação (71).

$$\hat{w}^*_{\tau+h|\tau,k} = g(\boldsymbol{\Delta}_{\tau+h|\tau}, \mathbf{b}) = logsig\left(\sum_{i=1}^{p}\beta_i . tanh\left(\sum_{j=1}^{N}\beta_{2p+(i-1)N+j}\delta_{\tau+h|\tau,j} + \beta_{p+i}\right) + \beta_0\right) \quad (71)$$

onde

$$logsig(u) = \frac{1}{1+e^{-u}} \quad (72)$$

A função logística (*logsig*) é limitada assintoticamente no intervalo (0,1), restringindo o espaço de busca e simplificando o algoritmo de treinamento. Por outro lado, o uso da *logsig* por si só não garante convexidade nos vetores de ponderação estimados, uma vez que não atua sobre a restrição "somar 1". Neste trabalho, a garantia de convexidade é dada pela reconciliação de pesos, em tempo de pós-processamento (seção 3.7).

Na condição de modelos não lineares complexos, todas as redes MLP têm seus parâmetros ajustados por otimização aproximada, i.e., sem solução exata. Nesta linha, um algoritmo bastante recomendado na prática é o *Back-propagation Levenberg-Marquardt* (Hagan & Menhaj, 1994; Mathworks, 2010b).

3.9. Política de treinamento

Para que se tenha robustez no ajuste dos parâmetros e hiperparâmetros dos modelos de rede neural, deve-se estabelecer uma política de treinamento estatisticamente adequada. Garantir a robustez estatística é uma prática particularmente importante em modelos de regressão não linear, onde o ajuste de parâmetros é feito por algoritmos de otimização aproximados, dependentes de valores iniciais (caso das redes neurais).

A política de treinamento usada neste trabalho é conhecida como *holdout* repetido (Witten & Frank, 2005): primeiro, uma porção da amostra de treinamento é separada, constituindo um novo conjunto chamado de conjunto de **validação**; depois, a porção que sobra do conjunto original – chamada de conjunto de **estimação** – é utilizada para treinar a rede neural diversas vezes, cada vez com uma configuração inicial diferente (ou seja, número de neurônios na camada escondida e pesos sinápticos iniciais diferentes). Para cada configuração treinada, a que apresentar melhor desempenho no conjunto de validação deve ser selecionada. Nas aplicações envolvendo séries temporais é uma boa prática a seleção dos pontos mais recentes da amostra de treinamento para composição do conjunto de validação; esta prática visa à manutenção da dependência serial no processo de aprendizado. Tipicamente, conjuntos de validação correspondem a 1/3 do conjunto de treinamento, sendo o restante da amostra utilizada para estimação.

No caso específico dos modelos NEW, uma questão prática a ser observada é que há duas formas de avaliar o desempenho de uma rede neural no conjunto de validação. A primeira, mais direta, é medir o quanto se erra na inferência das variáveis de saída, ou seja, na inferência dos pesos de combinação. Não deve-se confundir **pesos de combinação** (saídas dos modelos NEW) com **pesos sinápticos**, nome dado aos parâmetros intrínsecos das redes neurais (seção 3.8). A segunda forma de avaliar o desempenho na validação é, de posse dos vetores de pesos gerados, realizar a combinação dos previsores e avaliar o erro médio da **previsão combinada**. Esta segunda abordagem tem a vantagem de possuir interpretação mais direta; contudo, tendo havido um treinamento adequado da rede em questão, espera-se que a escolha entre um ou outro procedimento de avaliação não leve a resultados significativamente diferentes.

A Figura 14 lista, em pseudocódigo, a política de *holdout* repetido. Antes de passar-se à uma breve explicação da figura, é importante lembrar que dois aspectos principais caracterizam a configuração inicial de uma rede neural: (i) número de neurônios na camada escondida e (ii) pesos sinápticos iniciais.

Fixado um número de neurônios, cada nova inicialização de pesos sinápticos da rede é chamada de **replicação**[10]. Em cada replicação, o algoritmo de treinamento – caracterizado pela função TREINA na Figura 14 – é monitorado **época a época** (passo a passo), permitindo que se identifique – através da função TESTA – a configuração de rede com menor erro no conjunto de validação, mesmo que ela ocorra em uma época intermediária (inferior ao máximo de épocas permitidas); este esquema simula o que se chama de parada antecipada do treinamento (*early stopping*) (Bishop, 1995). Ao final de todos os laços (*loop*s) do pseudocódigo, tem-se em mãos a configuração de rede com melhor desempenho dentre todas as testadas.

[10] Em todos os experimentos de redes neurais neste trabalho foram consideradas 9 replicações, com teste de 1 a 30 neurônios em cada uma.

```
INÍCIO
  N:  número máximo de neurônios
  R:  número de replicações
  E:  número de épocas
  redeMin:  vetor 1..N de rede neural
  redeVal:  vetor 1..R de rede neural
  erroMin:  vetor 1..N de erro
  erroVal:  vetor 1..R de erro
  indMin:  índice de vetor
  rede:  rede neural
  erro:  erro
  tr:  conjunto de treinamento
  val:  conjunto de validação
  PARA neurônios = 1..N
     PARA replicação = 1..R
          rede = INICIA('mlp',neurônios)
          erroVal(replicação) = INFINITO
          PARA época = 1..E
               rede = TREINA(rede, tr)
               erro = TESTA(rede, val)
               SE error <  erroVal(replicação)
                    redeVal(replicação) = rede
                    erroVal(replicação) = erro
               FIM SE
          FIM PARA
     FIM PARA
     indMin = ÍNDICE(MÍNIMO(erroVal))
     redeMin(neurônios) = redeVal(indMin)
     erroMin(neurônios) = MÍNIMO(erroVal)
  FIM PARA
  indMin = ÍNDICE(MÍNIMO(erroMin))
  rede = redeMin(indMin)
  erro = erroMin(indMin)
  RETORNA rede, erro
FIM
```

Figura 14 – Política de treinamento (*hold-out* repetido). Para cada nova replicação, os parâmetros da rede neural são reiniciados e diferentes números de neurônios são testados. Ao final, a melhor configuração – aquela com menor erro médio no conjunto de validação – é selecionada.