5.2. Métodos baseados num modelo com tendência linear

5.2.0. Introdução

A idéia de amortecimento exponencial pode ser estendida para situações em que o nível da série não é constante, mas varia ao longo do tempo de acordo com uma tendência. Veremos nesta seção métodos desenvolvidos para modelos com tendência linear; na seção 5.3, veremos os métodos para modelos com tendência quadrática.

O modelo básico, como o da seção anterior, é dado pela combinação de um nível com um erro aleatório:

$$Z_t = \mu_t + \varepsilon_t$$

Agora, contudo, o nível não será mais uma constante, e sim uma função linear μ_t do tempo:

$$\mu_t = a + bt$$

Portanto, o *modelo linear* é dado por

$$Z_t = a + bt + \varepsilon_t \qquad (1)$$

onde a e b são constantes. Se o erro ε_t é uma sequência de variáveis aleatórias i.i.d., de média nula e variância σ^2, o valor esperado e a variância de Z_t serão dados por:

$$E(Z_t) = a + bt$$

As previsões para o instante $t = T+k$, feitas no instante atual $t = T$ por extrapolação da tendência b, são dadas por:

$$\hat{Z}_{T+k|T} = E(Z_{T+k} \mid \mathbf{Z}_T)$$

$$\hat{Z}_{T+k|T} = E[a + b(T+k) + \varepsilon_{T+k} \mid \mathbf{Z}_T] = \hat{a}_T + \hat{b}_T(T+k)$$

ou, para simplificar, chamando o instante atual de $T = 0$,

$$\hat{Z}_{T+k|T} = E\{a + bk + \varepsilon_{T+k} \mid \mathbf{Z}_T\} = \hat{a}_T + \hat{b}_T k$$

Em particular, a previsão um-passo-à-frente será dada por:

$$\hat{Z}_{T+1|T} = \hat{a}_T + \hat{b}_T \qquad (2)$$

onde $\hat{a}_T$ e $\hat{b}_T$ são as estimativas no nível e da tendência, respectivamente, feitas no instante atual $t = T$.

5.2.1. Método de amortecimento exponencial duplo de Brown

Este método estima os dois parâmetros a e b a partir de um sistema de duas equações envolvendo médias amostrais simples e duplas, como já tinha sido feito na Seção 4.2.2. Agora, porém, estas médias serão ponderadas exponencialmente, em vez de serem médias móveis ordinárias.

O método usa uma mesma constante de amortecimento α nas duas equações. A primeira equação é idêntica a do método de amortecimento exponencial simples (eq. 2, na seção 5.1.1):

$$M_T = \alpha Z_T + (1-\alpha)M_{T-1} \tag{3}$$

A segunda equação, envolvendo médias duplas, é:

$$M_T^{[2]} = \alpha M_T + (1-\alpha)M_{T-1}^{[2]} \tag{4}$$

Os valores das médias amortecidas simples e duplas são calculados recursivamente da forma acima, e depois inseridos nas equações:

$$\hat{a}_T \approx 2M_T - M_T^{[2]} \tag{5}$$

$$\hat{b}_T \approx \frac{\alpha}{1-\alpha}\left(M_T - M_T^{[2]}\right) \tag{6}$$

Note que esta eq. (6) é um pouco diferente daquela usada no método baseado em médias duplas ordinárias, não-ponderadas; (eq. 8, seção 4.2.2, reproduzida abaixo):

$$\hat{b}_T \approx \frac{2}{n-1}(M_T - M_T^{[2]}) \tag{7}$$

No método das médias móveis duplas ordinárias (eq. 7), o termo 2/(n-1) que multiplica a diferença entre as médias $M_T - M_T^{[2]}$ está expresso em termo do tamanho n da janela móvel usada; no método das médias móveis duplas amortecidas exponencialmente (eq. 6), o termo α/(1-α) está expresso em termos da constante de amortecimento α. Estas duas equações, contudo, podem ser consideradas equivalentes, em termos da *idade das observações* usadas para fazer as estimativas.

No Apênd. 1, vimos que, no método das médias móveis de ordem n, a média das idades I das observações usadas na estimação é dada por:

$$E(I) = \frac{n-1}{2}$$

No método de amortecimento exponencial simples, esta média é dada por:

$$E(I) = \frac{1-\alpha}{\alpha}$$

Portanto, para que estes dois métodos façam suas estimativas com base em observações que tenham em média a mesma idade, é preciso fazer que o n das médias móveis e o α do amortecimento exponencial obedeçam a relação

$$\frac{n-1}{2} = \frac{1-\alpha}{\alpha}$$

Note que os termos que multiplicam a diferença $M_T - M_T^{[2]}$ nas eqs. (6) e (7) são os recíprocos dos dois membros da igualdade acima; portanto, ao escrever a eq. (6), Brown garantiu que as médias amortecidas calculadas por seu método usem observações que tenham em média a mesma idade que as observações usadas no método das médias móveis ordinárias.

5.2.1.1. Escolha dos valores das constantes, e métodos de inicialização

Para começar as iterações do método de Brown, precisamos de um valor inicial para a constante de amortecimento α, e de valores iniciais adequados para as médias móveis simples e duplas.

A experiência sugere que os valores ótimos de α se encontram, na maioria das vezes, num intervalo restrito entre 0,1 e 0,2. Um valor α = 0,1 leva a uma previsão mais "conservadora", na qual a observação mais recente tem pouca influência; um valor α=0,2 leva a uma previsão que responde um pouco mais rápido a mudanças que ocorram na série. Na prática, o valor ótimo de α pode ser encontrado experimentalmente, por meio de uma busca em grade, ou por meio de um algoritmo de otimização não-linear (ver Seção 5.5).

Para obter os valores iniciais das médias simples e duplas, vários métodos têm sido sugeridos por diferentes autores. Quando a série disponível é suficientemente longa, uma possibilidade é separar a amostra em duas partes, e usar os dados da primeira parte para calcular estimativas que servirão de valores iniciais para as iterações do método de Brown nos dados da segunda parte. Algumas maneiras de fazer isto são:

Método 1:

Calculamos na primeira parte as séries de médias simples e duplas ordinárias (não-ponderadas), como feito na Seção 4.2.2. A seguir, usamos os valores destas médias obtidos no fim da primeira parte como valores iniciais para as iterações do método de Brown na segunda parte.

Por exemplo, se usamos médias móveis de ordem n = 3, os passos são:

- calcular as três primeiras médias móveis:

$$MM_3 = \frac{Z_3 + Z_2 + Z_1}{3} \qquad MM_4 = \frac{Z_4 + Z_3 + Z_2}{3} \qquad MM_5 = \frac{Z_5 + Z_4 + Z_3}{3}$$

- a partir destas, calcular a primeira média móvel dupla:

$$MM_5^{[2]} = \frac{MM_5 + MM_4 + MM_3}{3}$$

As primeiras estimativas de nível e tendência serão portanto $\hat{a}_5$ e $\hat{b}_5$, feitas com base em M_5 e em , usando as eqs. (**3**) e (**4**). As médias móveis do instante de t = 6 em diante serão feitas por amortecimento exponencial, usando as eqs. (**1**) e (**2**).

Método 2:

Ajustamos um modelo de regressão linear sobre os primeiros m valores da série, e obtemos a partir daí estimativas iniciais para o nível a_0 e a declividade b_0:

$$Z_t = a_0 + b_0 t$$

A partir destes coeficientes a_0 e b_0, podemos obter valores iniciais das médias, por meio das eqs (**3**) e (**4**):

$$\hat{a}_T \approx 2M_T - M_T^{[2]} \qquad \text{donde } M_T^{[2]} = 2M_T - \hat{a}_T \qquad \text{(i)}$$

$$\hat{b}_T \approx \frac{\alpha}{1-\alpha}\left(M_T - M_T^{[2]}\right) \qquad \text{donde } M_T^{[2]} = M_T - \frac{1-\alpha}{\alpha}\hat{b}_T \qquad \text{(ii)}$$

Igualando as equações (i) e (ii), obtemos

$$2M_T - \hat{a}_T = M_T - \frac{1-\alpha}{\alpha}\hat{b}_T \qquad \rightarrow \qquad M_T = \hat{a}_T - \left(\frac{1-\alpha}{\alpha}\right)\hat{b}_T \qquad \text{(iii)}$$

Inserindo o valor de M_T em (iii) na eq. (ii),

$$M_T^{[2]} = \hat{a}_T - \left(\frac{1-\alpha}{\alpha}\right)\hat{b}_T - \frac{1-\alpha}{\alpha}\hat{b}_T \quad \rightarrow \quad M_T^{[2]} = \hat{a}_T - 2\left(\frac{1-\alpha}{\alpha}\right)\hat{b}_T \qquad \text{(iv)}$$

Os valores iniciais das médias móveis serão portanto, usando (iii) e (iv):

$$M_0 = \hat{a}_0 - \left(\frac{1-\alpha}{\alpha}\right)\hat{b}_0$$

$$M_0^{[2]} = \hat{a}_0 - 2\left(\frac{1-\alpha}{\alpha}\right)\hat{b}_0$$

As estimativas de a e b e as previsões de $\hat{Z}_t$ serão feitas a partir do instante $t = 1$. Note, contudo, que os valores estimados de $\hat{Z}_1$ a $\hat{Z}_m$ não serão realmente "previsões", já que estes m primeiros valores já são todos conhecidos, e foram usados no ajuste do modelo de regressão linear. No cálculo dos erros e avaliação da acurácia do método, é portanto mais sensato considerar apenas as previsões a partir de $t = m+1$.

Método 3:

Se a série disponível é curta, e não queremos separar dados para inicialização, valores iniciais aproximados podem ser obtidos fazendo-se:

- médias móveis iniciais iguais ao primeiro valor observado

$$M_1 = M_1^{[2]} = Z_1$$

- nível inicial: igual ao primeiro valor observado

$$\hat{a}_1 = Z_1$$

- tendência inicial: igual à diferença entre os dois primeiros valores observados:

$$\hat{b}_1 = Z_2 - Z_1$$

ou à média da diferenças entre os k primeiros valores observados; por exemplo:

$$\hat{b}_1 = \frac{Z_4 - Z_1}{3}$$

ou ainda:

$$\hat{b}_1 = \frac{(Z_2 - Z_1) + (Z_4 - Z_3)}{2}$$

5.2.1.2. Exemplos

A Fig. **1** mostra exemplos da aplicação do método de Brown na previsão um-passo-à-frente da série *nhtemp* (temperatura media anual em New Haven, Connecticut, em graus ^{o}F, de 1912 a 1971; parte do pacote *datasets*). Os valores iniciais foram determinados pelo *Método 3* acima. O valor ótimo da constante de amortecimento, encontrado pela função de otimização `optim`, foi α=0,076. A figura compara as previsões obtidas com este valor com as previsões feitas usando dois outros valores. É possível observar que valores pequenos de α (como, no exemplo, 0.020) levam a uma série de previsões muito amortecida, que reage lentamente às variações ocorridas na série. Por outro lado, valores de α muito altos (como, no exemplo, 0.200) levam a uma série de previsões que reage rápido demais às variações da série, e podem por isso ser indevidamente afetadas pelo ruído presente nos dados. O valor encontrado por otimização levou a uma série previsões intermediária entre estes dois extremos, e resultou por isso no menor erro quadrático (MSE).

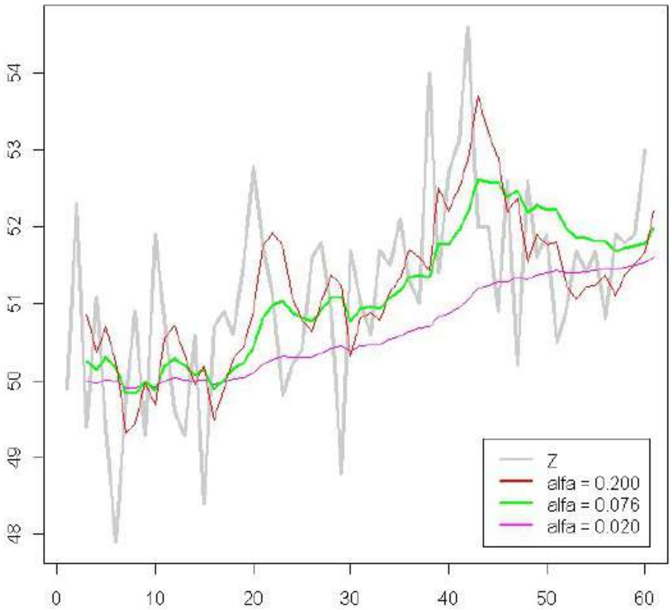


Figura 1. Exemplos da aplicação do método de Brown, série *nhtemp*

5.2.1.3. Conclusão

O método linear de Brown é um dos métodos que podem ser usados para a previsão de séries não-estacionárias e não-sazonais. No passado, segundo [1], costumava ser o método preferido, por exigir menos cálculos do que o método das médias móveis duplas (seção 4.2.2), e ser mais fácil de implementar, por usar apenas uma constante de amortecimento, do que o método de Holt (seção 5.3). Atualmente, porém, o número de constantes ou dos cálculos necessários para a previsão em geral não é mais um empecilho à utilização destes métodos; isto já foi observado numa edição posterior do mesmo livro citado acima [2]. Não há, por isso, critérios rígidos para decidir qual dos três métodos utilizar, num determinado problema; a escolha vai ser feita levando em conta a conveniência do usuário (por exemplo, qual destes métodos já está implementado no pacote estatístico que você irá usar?), ou o resultado de testes empíricos feitos nos dados disponíveis.

5.2.2. Método de amortecimento exponencial de Holt

Neste método, os parâmetros – o nível a e a tendência b – são atualizados diretamente de forma recursiva, sem requerer o uso de médias móveis. São usadas duas equações de atualização, uma para cada parâmetro:

$$\hat{a}_T = \alpha Z_T + (1-\alpha)(\hat{a}_{T-1} + \hat{b}_{T-1}) \qquad (8)$$

$$\hat{b}_T = \beta(\hat{a}_T - \hat{a}_{T-1}) + (1-\beta)\hat{b}_{T-1} \qquad (9)$$

Estas equações podem parecer complicadas, mas o raciocínio por trás delas é simples. Lembre-se de que a idéia básica de qualquer método de amortecimento exponencial é a de obter a cada instante estimativas atualizadas, combinando uma estimativa obtida a partir da informação mais recente com uma estimativa obtida a partir de informação que já era conhecida no instante anterior (Fig. 2).

$$\begin{bmatrix}\text{estimativa}\\\text{final}\end{bmatrix} = \alpha \begin{bmatrix}\text{estimativa}\\\text{calculada a}\\\text{partir da}\\\text{informação}\\\text{mais recente}\end{bmatrix} + (1-\alpha) \begin{bmatrix}\text{estimativa}\\\text{calculada a}\\\text{partir da}\\\text{informação}\\\text{disponível}\\\text{no instante}\\\text{anterior}\end{bmatrix}$$

Figura 2. Idéia básica da estimação por amortecimento exponencial

Suponha que estamos no instante $t = T-1$, e temos as estimativas $\hat{a}_{T-1}$ e $\hat{b}_{T-1}$. A partir delas, podemos fazer uma estimativa para o nível no instante T:

$$\hat{a}_{T|T-1} = \hat{a}_{T-1} + \hat{b}_{T-1}$$

O sub-índice "T|T−1" indica que se trata de uma estimativa que o parâmetro deverá assumir no instante T, feita com base na informação disponível até o instante $T-1$.

Suponha agora que estamos no instante T, e observamos o valor Z_T. Uma nova estimativa do nível, feita com base nesta informação, pode ser obtida fazendo simplesmente:

$$\hat{a}_{T|T} = Z_T$$

A equação (8) combina linearmente estas duas estimativas, usando o peso α:

$$\hat{a}_T = \alpha \hat{a}_{T|T} + (1-\alpha)\hat{a}_{T|T-1}$$

$$\hat{a}_T = \alpha Z_T + (1-\alpha)(\hat{a}_{T-1} + \hat{b}_{T-1})$$

Da mesma forma, a eq. (9) calcula uma estimativa atualizada da tendência b pela combinação entre uma estimativa que poderia ser feita a partir da informação disponível até o instante $T-1$ (uma estimativa *naïve*, dada pela simples repetição da estimativa $\hat{b}_{T-1}$, a melhor disponível até aquele instante) com uma estimativa de b dada pela diferença entre a estimativa $\hat{a}_T$ do nível feita no instante atual (pela eq. 8, com base na informação mais recente, trazida por Z_T) e a estimativa $\hat{a}_{T-1}$ do nível feita no instante anterior:

$$\hat{b}_T = \beta(\hat{a}_T - \hat{a}_{T-1}) + (1-\beta)\hat{b}_{T-1}$$

5.2.2.1. Escolha dos valores das constantes α e β, e métodos de inicialização

Para iniciar as iterações do método de Holt, precisamos de valores para as constantes de amortecimento α e β, e de valores iniciais adequados para os parâmetros a e b. Os valores ótimos de α e β podem ser encontrados empiricamente, por meio de uma busca em grade, ou por meio de um algoritmo de otimização não-linear (ver Seção 5.5).

Para obter os valores iniciais dos parâmetros a e b, vários métodos diferentes são sugeridos na literatura. Quando a série é longa, uma possibilidade é particioná-la em duas partes, e usar os dados da primeira parte para calcular estimativas $\hat{a}$ e $\hat{b}$ por meio de regressão linear; estas estimativas servirão de valores iniciais para as iterações feitas nos dados da segunda parte.

Outra possibilidade é estimar os valores iniciais por *backcasting* ("previsão para trás") - ou seja, revertendo a serie, e estimando nível e tendência iterativamente do último valor observado até o primeiro. Os valores $\hat{a}_1$ e $\hat{b}_1$ estimados para o primeiro instante, ao final das iterações retroativas, serão então usados como valores iniciais para as previsões feitas na ordem correta, para a frente. Esta técnica é mais comumente usada, contudo, nos modelos ARIMA, que veremos depois (Cap. **6**).

Se a série disponível for curta, e não queremos separar dados para inicialização, valores iniciais aproximados podem ser obtidos destas formas simplificadas:

- para o nível inicial: usando a primeira observação
$$\hat{a}_1 = Z_1 \tag{i}$$
- para a tendência inicial: usando uma destas opções:
  - a diferença entre as duas primeiras observações:
  $$\hat{b}_1 = Z_2 - Z_1 \tag{ii}$$
  - a diferença entre dois pares de observações consecutivas, como em:
  $$\hat{b}_1 = \frac{(Z_2 - Z_1) + (Z_4 - Z_3)}{2} \tag{iii}$$
  - a média das diferenças entre as *k* primeiras observações; por exemplo
  $$\hat{b}_1 = \frac{Z_4 - Z_1}{3} \tag{iv}$$

5.2.2.2. Trend dampening (amortecimento da tendência)

Métodos de previsão, especialmente para horizontes longos, tendem a superestimar a tendência. Na prática, nenhuma variável pode crescer (ou decrescer) indefinidamente; em algum momento, a tendência irá se anular (o nível se tornará constante), ou irá se reverter (a curva mudará de direção). Isto é bem conhecido na Economia, que prefere distinguir *tendência* de *ciclo* (ver Cap. **3**).

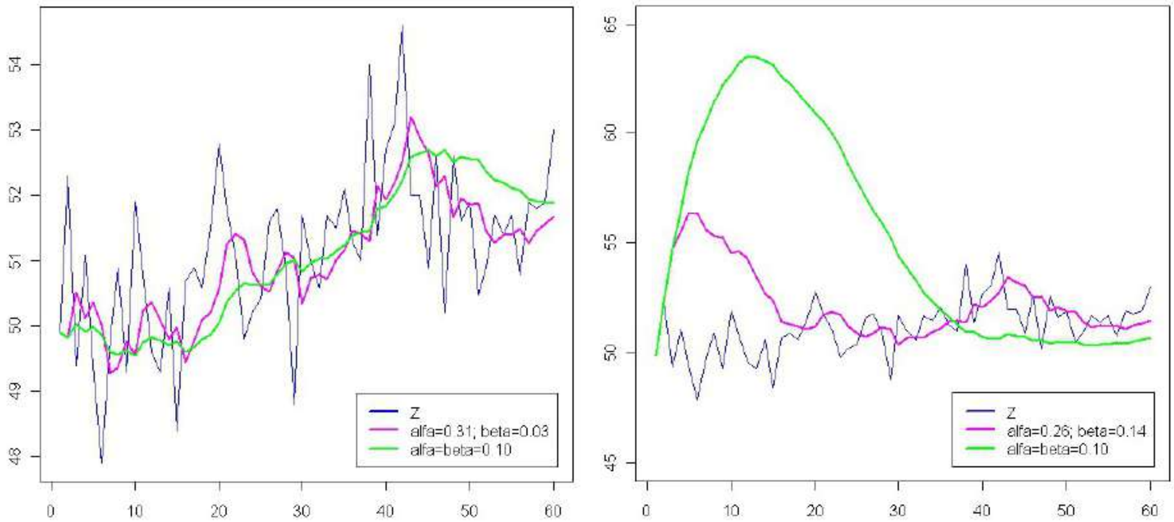
As técnicas de amortecimento da tendência (*trend dampening*) buscam reduzir o valor da tendência progressivamente, de forma proporcional à magnitude do horizonte de previsão; em geral, para horizontes longos, o nível se aproxima assintoticamente de uma constante. Uma forma de conseguir isto é usar uma constante de amortecimento da tendência, $0 \le \varphi \le 1$, e fazer a previsão com a tendência corrigida como na eq. (**10**):

$$\hat{Z}_{T+k|T} = \hat{a}_T + \hat{b}_T \sum_{i=1}^{k} \varphi^i \tag{10}$$

5.2.2.3. Exemplos

O gráfico da Fig. **3A** mostra a série *nhtemp* (a mesma usada na Fig. **1**) e as previsões feitas pelo método de Holt usando os valores das constantes de amortecimento que minimizam o MSE (em magenta), encontrados pela função `optim`, e um par de outros valores escolhido para comparação. Para obter os valores iniciais dos parâmetros, usamos o método (i) da seção anterior para o nível $\hat{a}_1$ e o método (iv) para a tendência $\hat{b}_1$, com k=9.

O gráfico da Fig. **3B** mostra a mesma série, e serve para ilustrar o efeito das escolha inadequada dos valores iniciais. Novamente, usamos os valores ótimos das constantes de amortecimento, e um par de outros valores para comparação. Contudo, para obter a estimativa inicial da tendência $\hat{b}_1$ usamos agora o método (ii), que se mostrou claramente inadequado. Note que o valor ótimo de β encontrado pelo `optim` é muito maior do que o da figura A, indicando que foi necessário fazer fortes alterações nas estimativas da tendência, ao longo das iterações, para corrigir a má estimativa usada como valor inicial.



A - Previsão com constantes otimizadas (magenta) B – Efeito de valores iniciais inadequados

Figura 3. Exemplos da aplicação do método de Holt, série *nhtemp*

O gráfico da Fig. 4 mostra a mesma série, e compara previsões feitas pelos métodos de Brown e de Holt, ambos usando os valores das constantes de amortecimento que minimizam o MSE, e pelo método de Médias Móveis Duplas com n=8 (valor escolhido numa busca no intervalo de n=4 a n=10).

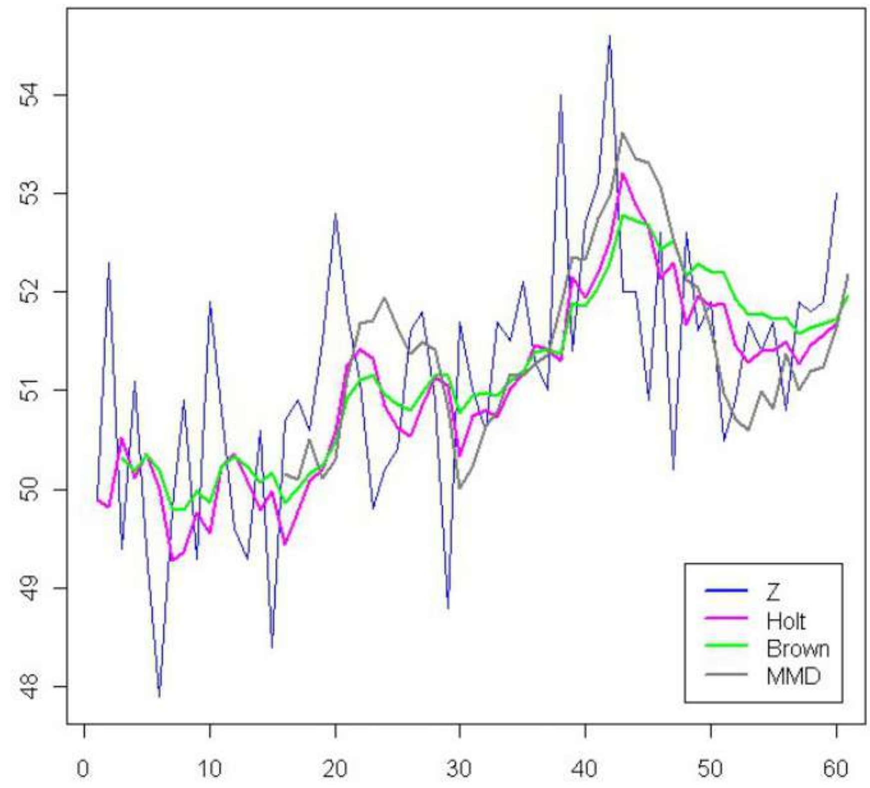


Figura 3. Aplicação dos métodos de Holt, Brown e Médias Móveis Duplas, série *nhtemp*

As previsões feitas pelos métodos de Holt e de Brown parecem ser razoavelmente similares, enquanto que as feitas por MMD parecem ter maiores erros. A Tab. 1 mostra os valores do MSE e do MAPE dos três métodos, calculados no intervalo em que foi feita a otimização dos parâmetros (t = 20:60). O método de Holt obteve o melhor resultado em termos do MSE, mas o de Brown foi o melhor em termos do MAPE (note porém que as constantes foram otimizadas para minimizar o MSE, nos dois métodos, não o MAPE).

Tabela 1. Comparação dos erros dos três métodos
calculados no intervalo de t=20 a t=60

	MSE	MAPE
Holt	1,206	1,68
Brown	1,215	1,63
Médias Móveis Duplas	1,368	1,83

5.2.2.4. Conclusão

O método de Holt provavelmente é o mais usado atualmente para a previsão de séries não-estacionárias e não-sazonais, já que pode ser considerado como um caso particular do método de Holt-Winters (seção 5.4), implementado em vários programas estatísticos ou linguagens de programação (no R, por exemplo, há a função `HoltWinters`, a partir da qual podem ser calculadas as previsões por AES, Holt, e HW).

Em termos de acurácia das previsões, contudo, os métodos de Holt e de Brown tendem a obter resultados bastante similares, desde que seja feita uma escolha adequada dos valores iniciais para as estimativas dos parâmetros a e b.

Referências

[1] Makridakis, S.; Wheelwright, S. C.; Mcgee, R. J. (1983). *Forecasting - Methods and applications*. 2nd ed. John Wiley & Sons.

[2] Makridakis, S., Wheelwright S.C., Hyndman, R. J. (1998) *Forecasting – Methods and Applications*. 3rd ed. New York: John Wiley & Sons.