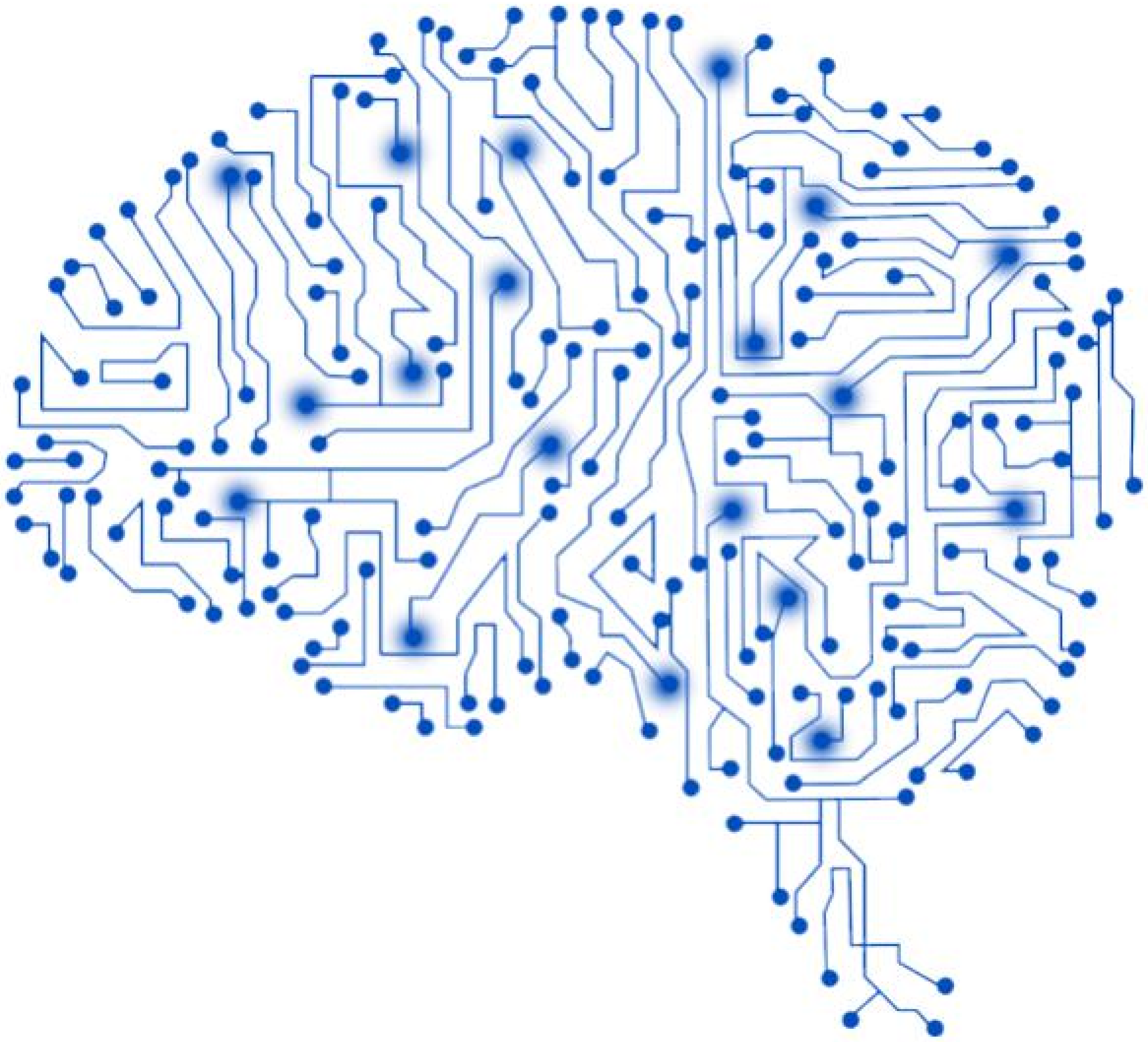
Cultura de Dados



Introdução a Análise de Séries Temporais

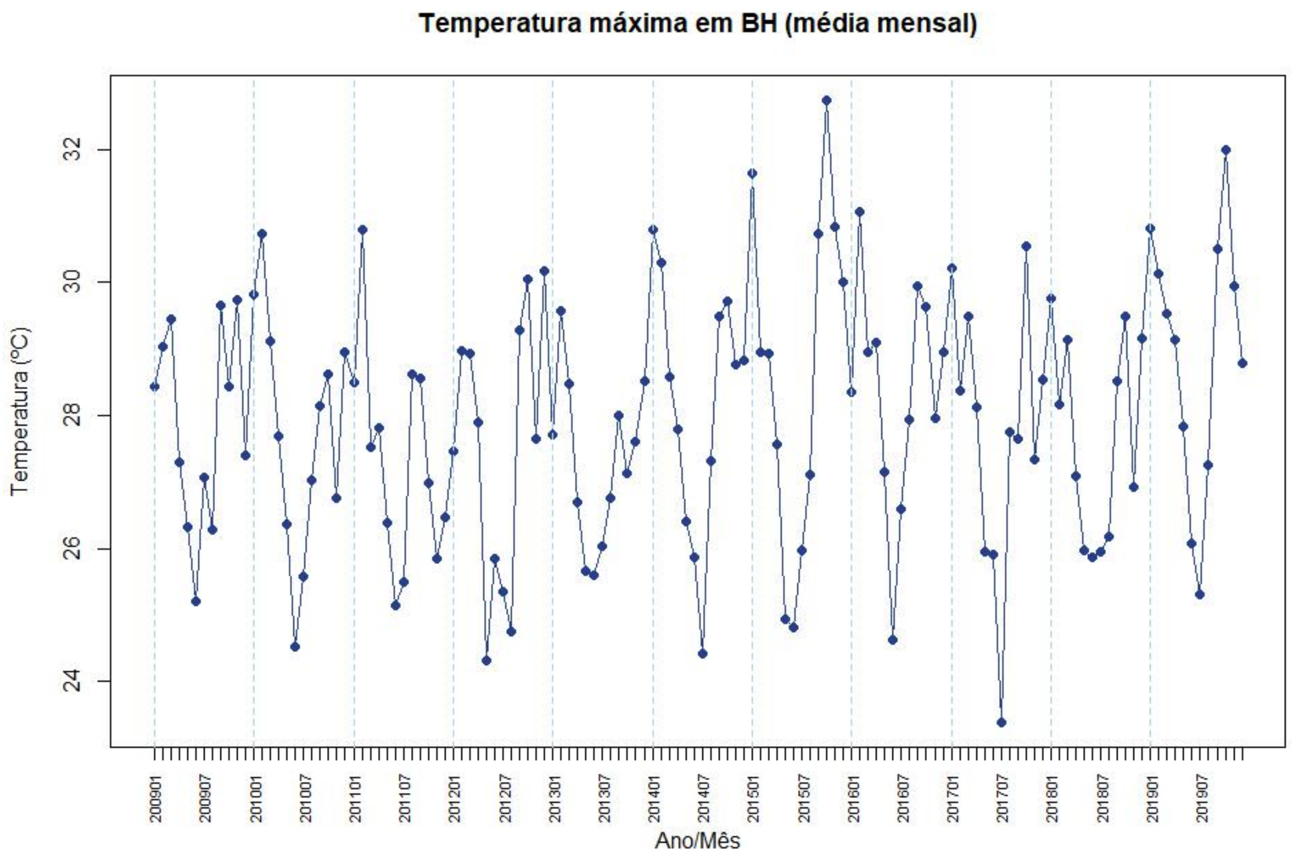
Kelver Campanha - Jan/2020

Introdução

Definição: De forma genérica, uma série temporal é um conjunto de observações obtidas sequencialmente no tempo (Cotação de uma determinada ação, temperatura, etc.).

Característica: A variável possui uma relação de dependência com seu(s) estado(s) anterior(es).

Objetivo: Entender e modelar a relação de dependência entre os estados da variável e fazer previsões para o futuro.



dados disponíveis em https://www.agritempo.gov.br/

Introdução

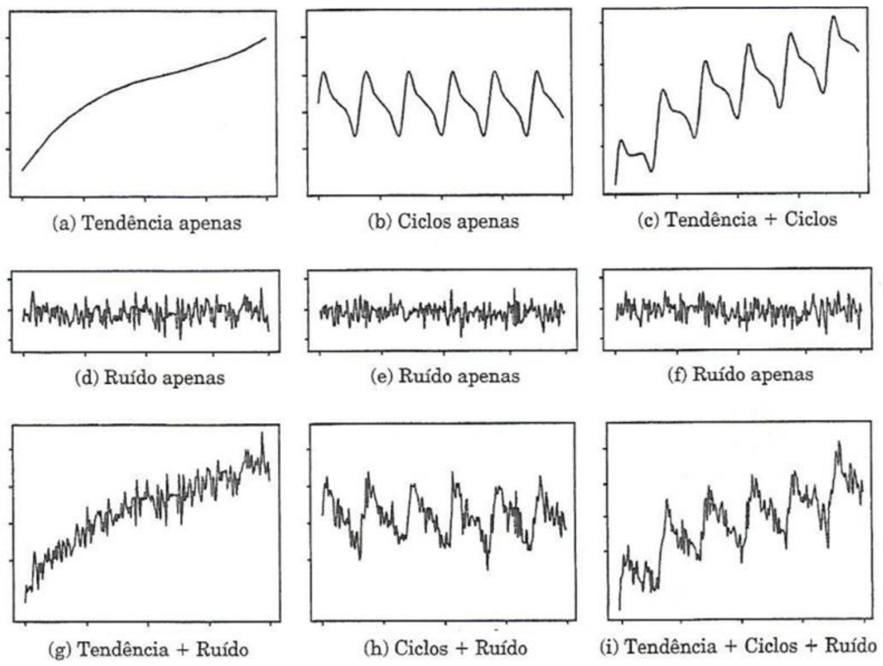
A natureza dinâmica de séries temporais faz com que modelos tenham que ter adaptabilidade no tempo, de tal modo que o princípio da parcimônia se faz muito presente no estudo de séries temporais, evitando a sobrefixação de parâmetros.

Componentes de uma série temporal:

- Ruído
- Tendência/Nível
- Ciclo/Sazonalidade

Tipos de Séries Temporais:

- Discreta
- Contínua
- Multivariada



Medotologia

Uma das metodologias mais utilizadas para a análise de séries temporais é a de Box e Jenkins (1976). Essa modelagem implica no ajuste de modelos autorregressivos integrados de médias móveis em um conjunto de dados ordenados no tempo.

A modelagem proposta por Box & Jenkins é da forma:

$$Y_t = \mu + \sum_{k=0}^{\infty} \Psi_k u_{t-k} = \mu + \Psi(B) u_t$$

onde o filtro linear Ψ é definido por:

$$\Psi(B) = \frac{\theta(B)}{\phi(B)}$$

onde $\theta(B)=1-\theta_1 B-\theta_2 B^2-...-\theta_q B^q$ e $\phi(B)=1-\phi_1 B-\phi_2 B^2-...-\phi_p B^p$ são polinômios de graus q e p, respectivamente. Assim temos:

$$\tilde{Y}_t = Y_t - \mu$$

$$\tilde{Y}_t = \Psi(B) u_t$$

Medotologia

Desta forma, os modelos de Box & Jenkins são dados por:

$$\phi(B)\widetilde{Y}_t = \theta(B)u_t$$

onde μ_t é um ruído branco, geralmente Gaussiano. De acordo com Box & Jenkins, o modelo acima é denominado ARMA(p,q), e dele pode-se escrever:

$$\widetilde{Y}_t = \theta(B)\phi^{-1}(B)\mu_t$$

Médias Móveis (MA)

- O modelo que tem $\theta(B)=1$ é chamado Modelo Médias Móveis. Notação: MA(q)
- O nome Médias Móveis vem do fato que Y_t é uma função soma algébrica ponderada dos μ_t que se movem no tempo.

Auto Regressivos (AR)

- O modelo que tem $\phi(B)=1$ é chamado Modelo Auto-Regressivo. Notação: AR(p)
- O nome Auto-Regressivo se deve ao fato de que Y_t no instante t é função dos Y's nos instantes anteriores a t.

AR-MA (ARMA)

- É o modelo que tem tanto uma parte AR ($\phi(B)\neq1$) como uma parte MA ($\theta(B)\neq1$). Notação: ARMA(p,q)

Medotologia

	Médias Móveis (MA)	Auto Regressivos (AR)	AR-MA (ARMA)
Estacionariedade	Todo modelo MA é estacionário.	Os modelos AR devem ter as raízes do polinômio $\Psi^{-1}(B) = \phi(B) = 0$ fora do círculo unitário como condição de estacionariedade.	Um modelo ARMA(p,q) será estacionário se $\phi(B)$ satisfizer as condições de estacionariedade de um modelo AR.
Inversibilidade	Os modelos MA devem ter as raízes do polinômio $\pi^{-1}(B) = \theta(B) = 0$ fora do círculo unitário como condição de inversibilidade.	Todo modelo AR é inversível.	Um modelo ARMA(p,q) será inversível se $\theta(B)$ satisfizer as condições de inversibilidade de um modelo MA.

Medotologia

Um processo não estacionário homogêneo é dado por

$$\nabla Y_t = (1 - B)Y_t = w_t$$

onde ▼ é um polinômio do tipo AR. Se $w_t = ▼^d Y_t$ é estacionária, podemos representar w_t por um modelo ARMA(p,q), ou seja,

$$\phi(B) w_t = \theta(B) \mu_t$$

Dizemos que Y_t segue um modelo Auto-Regressivo - Integrado - Média Móvel, ou ARIMA(p,d,q).

Seja ξ(B) o operador auto-regressivo não estacionário de ordem p+d:

$$\xi(B) Y_t = \theta(B) \mu_t$$

onde

$$\xi(B) = \phi(B) \nabla^d = \phi(B)(1 - B)^d$$

Portanto o modelo ARIMA supõe que a d-ésima diferença da série Y_t pode ser representada por um modelo ARMA, estacionário e inversível. Na maioria dos casos usuais, d = 1 ou 2.

Medotologia

Seja w_t uma série temporal estacionária. A função de autocovariância para w_t é dada por

$$\gamma_k = Cov(w_t, w_{t-k}) = E([w_t - E(w_t)][w_{t-k} - E(w_{t-k})])$$

Assumindo que $E(w_t)=E(w_{t-k})=0$, temos

$$\gamma_k = E([w_t][w_{t-k}])$$

A função de autocorrelação (FAC) pode ser então ser expressa como

$$\rho_k = \frac{\gamma_k}{\gamma_0}, k = 0,1,2,\ldots$$

onde $\gamma_0=Var(w_t)$.

A idéia de autocorrelação pode ser estendida. Se medirmos a correlação entre duas observações seriais, Y_t e Y_{t+k} , eliminando a dependência dos termos intermediários, temos o que se denomina autocorrelação parcial (FACP), representada por:

$$Cov(Y_t, Y_{t-k} \mid Y_{t-1}, \ldots, Y_{t-(k+1)})$$

Medotologia

Através das equações de Yule-Walker e utilizando a Regra de Cramer, a FACP pode ser generalisada como

$$\phi_{kk} = \frac{|P_k^*|}{|P_k|}$$

onde | . | é o determinante da matriz, P_k é a matriz de autocorrelação e P_k^* é a matriz P_k com a última coluna substituída pelo vetor de autocorrelações.

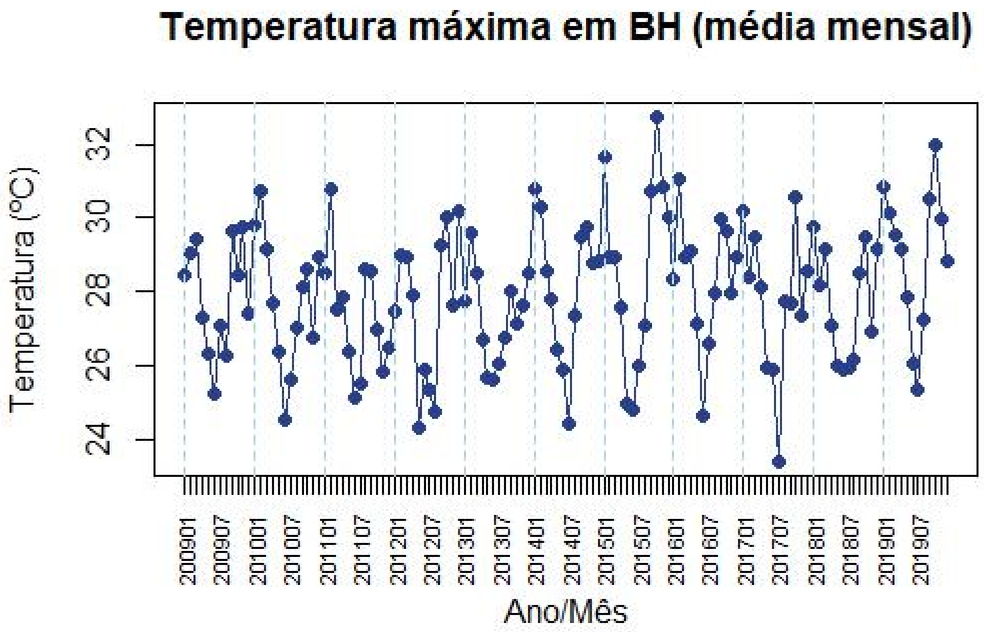
A estimação de Parâmetros do modelo de séries temporais pode ser feita por mínimos quadrados, pelo método do momentos ou pelo método de máxima verossimilhança. Neste material não será abordado mais detalhes, mas são os mesmos procedimentos tal como utilizados nos ajustes clássicos de regressão.

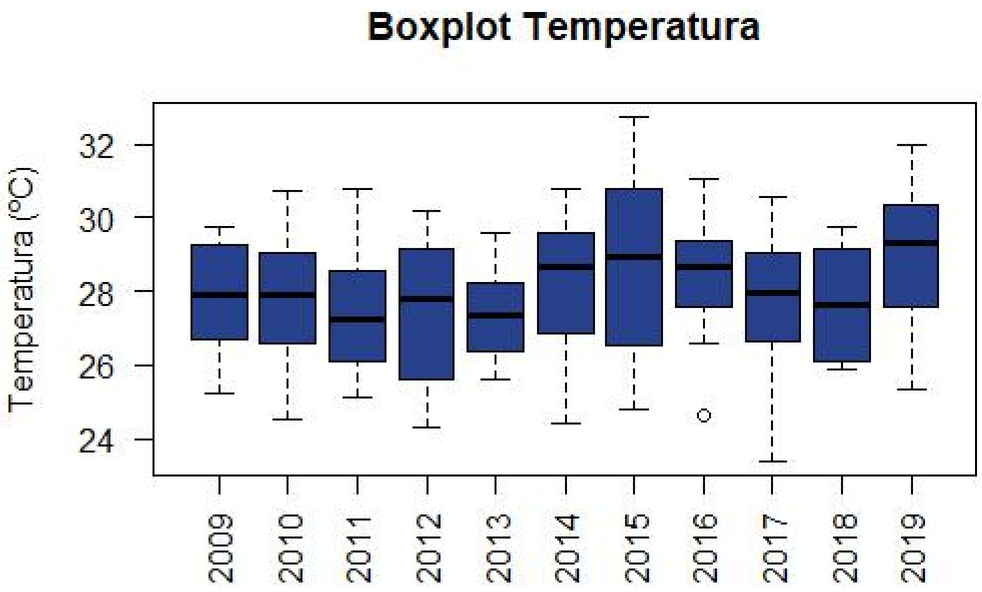
Quando observa-se a tendência do processo em repetir um certo tipo de comportamento dentro de um período sazonal (a cada 12 meses em séries mensais, por exemplo), é necessário a adição de um parâmetro 'S' aos modelos.

Na sequência, vejamos mais detalhes com exemplos práticos.

Desenvolvimento

Vontando a série temporal das temperaturas na cidade de Belo Horizonte, vamos inicialmente avaliar descritivamente os dados.

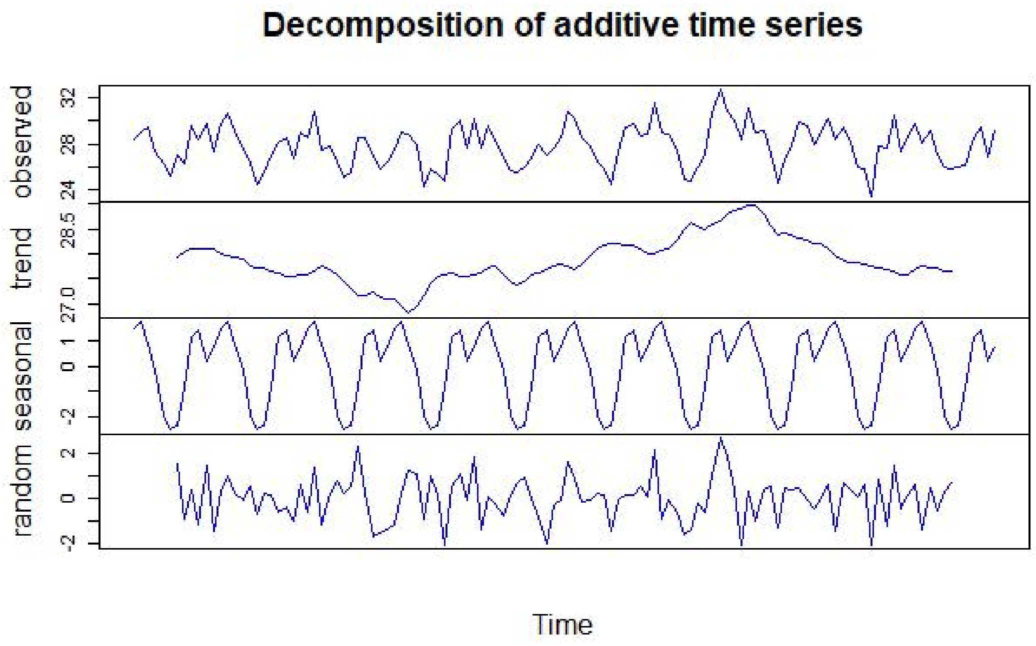


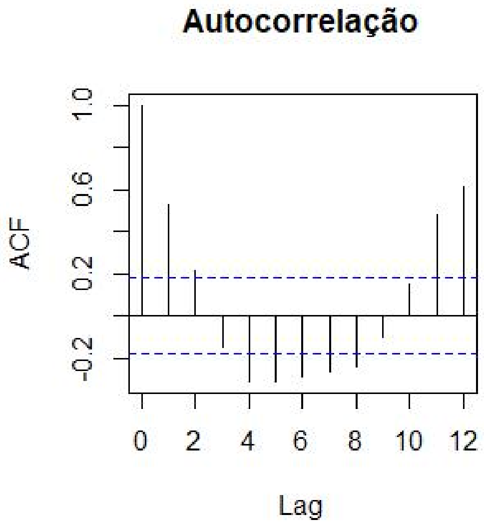


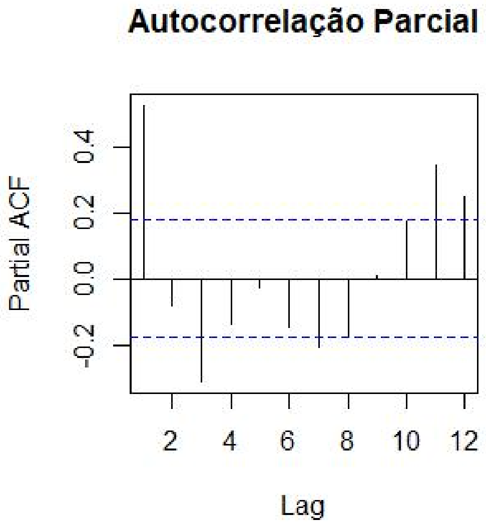
- Teste para a raiz unitária indica estacionaridade.
- Média mensal para as temperaturas máximas possui mediana próxima de 28º e distribuição aproximadamente simétrica.

Desenvolvimento

A decomposição da série temporal, tal como a FAC e FACP não dão insumos para a paramtrização do modelo. A função auto-arima pode ser utlizada para o tuning dos parâmetros (p,d,q)(P,D,Q).

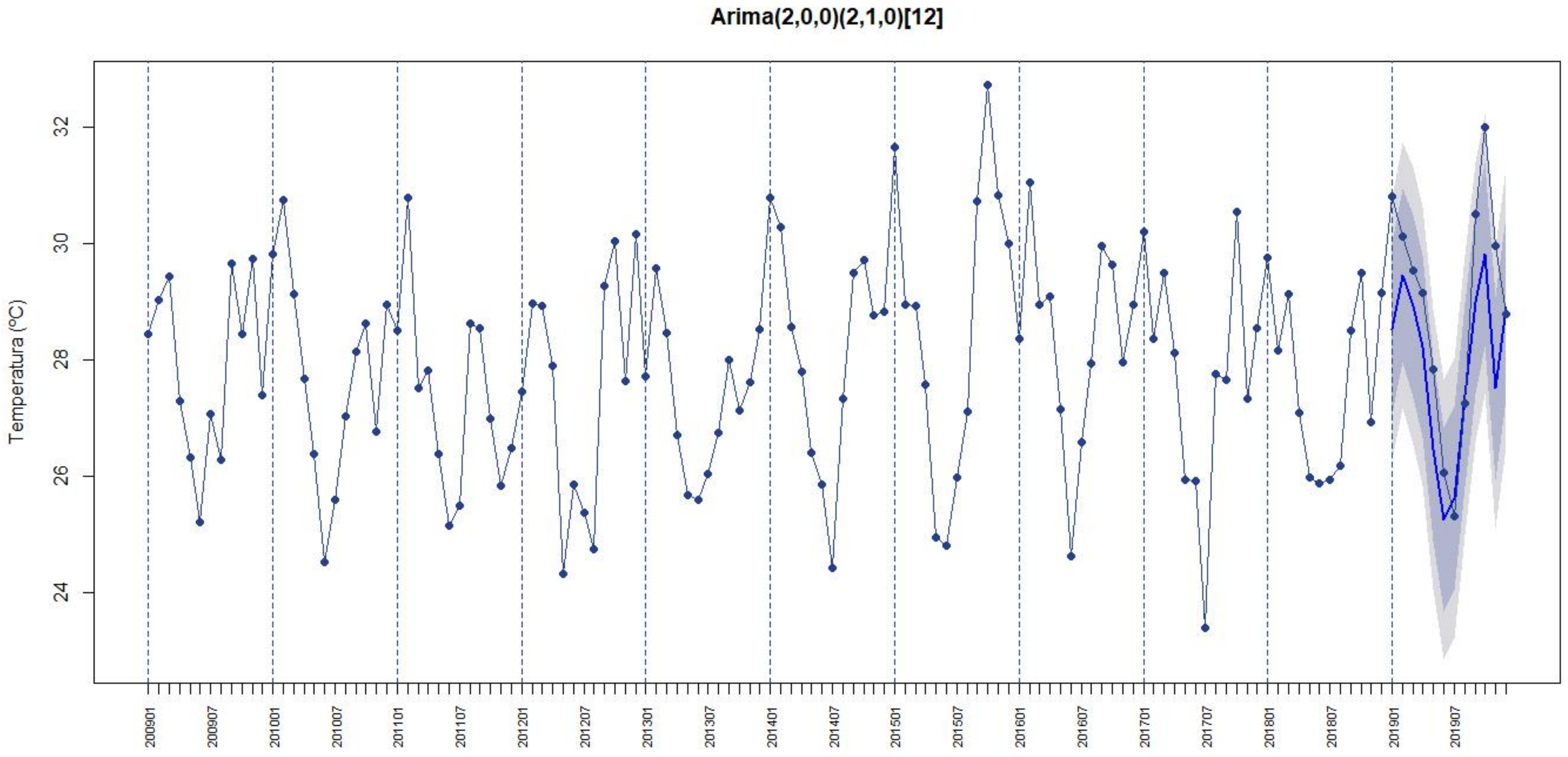






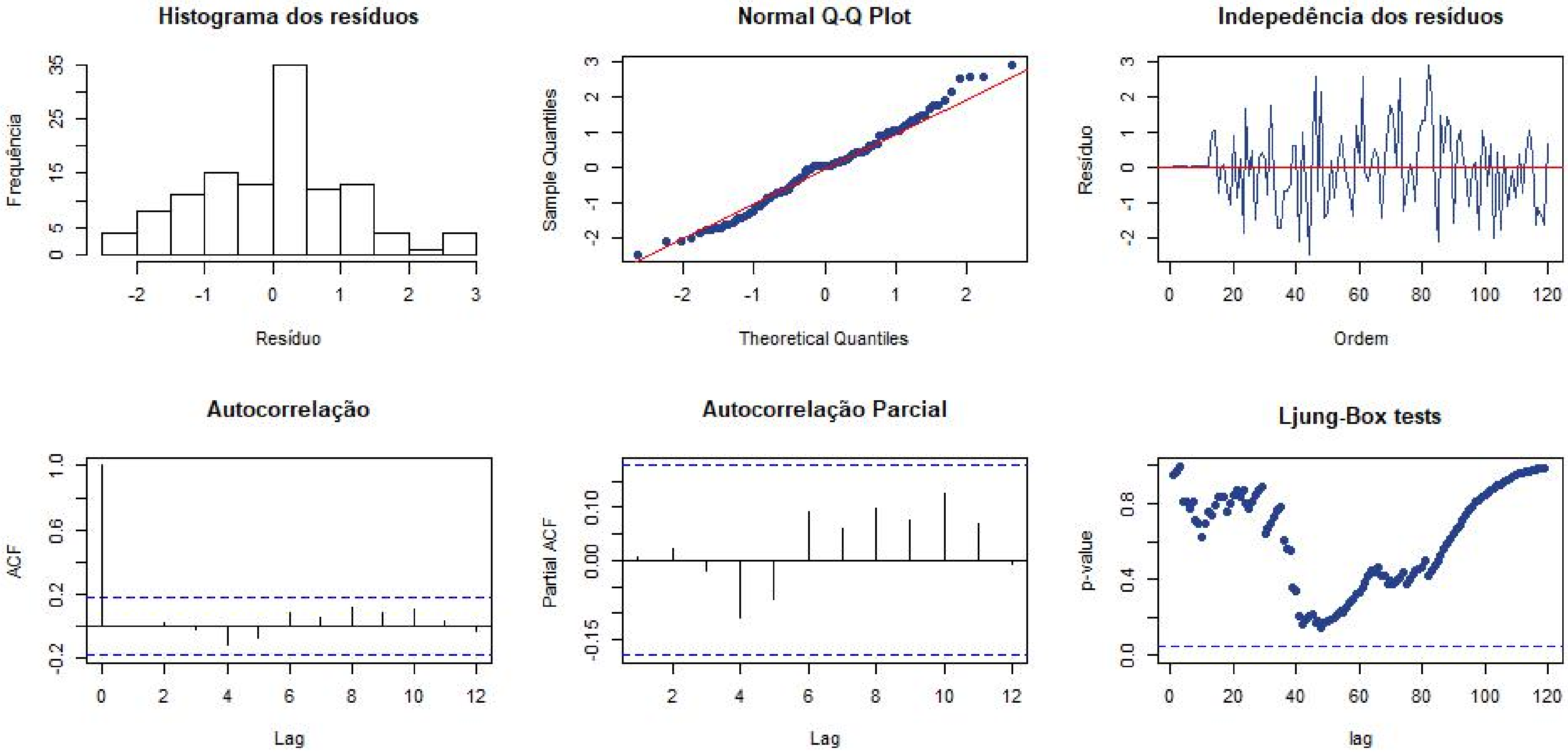
Desenvolvimento

As previsões para a média da temperatura máxima mensal em BH no ano de 2019 é apresentada a seguir, tal como seus respectivos intervalos de confiança.



Desenvolvimento

A análise de resíduos não nos revela nenhuma violação aos pressupostos do modelo. O ajuste esta, portanto, apto para novas previsões.

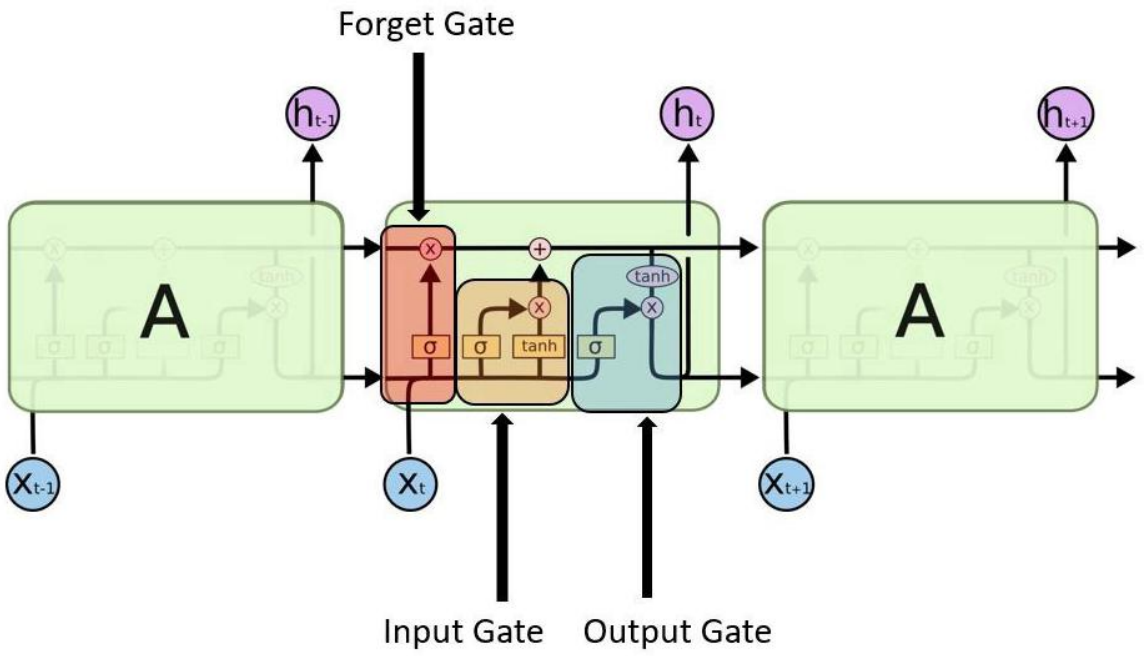


Desenvolvimento

A memória de longo prazo (LSTM) é uma arquitetura artificial de rede neural recorrente (RNN).

Uma unidade LSTM comum é composta por uma célula, um portão de entrada, um portão de saída e um portão de esquecer. A célula lembra valores em intervalos de tempo arbitrários e os três portões regulam o fluxo de informações para dentro e para fora da célula.

As redes LSTM são adequadas para classificar, processar e fazer previsões com base em dados de séries temporais, pois pode haver atrasos de duração desconhecida entre eventos importantes em uma série temporal.



Referências

[1] Notas de aula dos professores Thiago Rezende e Glaura Franco (DEST-UFMG).
[2] Material do curso: Deep Learning com Python de A a Z: O Curso Completo (Udemy).
[3] https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e
[4] https://www.monolitonimbus.com.br/machine-learning-de-series-temporais-lstm/
[5] https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b
[6] https://www.geeksforgeeks.org/python-arima-model-for-time-series-forecasting/