COMUNICAÇÃO TÉCNICA

Nº 177315

Detecção de anomalias, interpolação e previsão em tempo real de séries temporais para operação de reservatórios e distribuição de água

Leonardo Fonseca Larrubia
Chang Chiann
Olga Satomi Yoshida

Palestra apresentada no WORKSHOP TRM TECNOLOGIAS REGULATÓRIAS E METROLÓGICAS, 4., 2021., São Paulo. 16 slides

Instituto de Pesquisas Tecnológicas do Estado de São Paulo
S/A - IPT
Av. Prof. Almeida Prado, 532 | Cidade Universitária ou
Caixa Postal 0141 | CEP 01064-970
São Paulo | SP | Brasil | CEP 05508-901
Tel 11 3767 4374/4000 | Fax 11 3767-4099

www.ipt.br

Detecção de anomalias, interpolação e previsão em tempo real de séries temporais para operação de reservatórios e distribuição de água

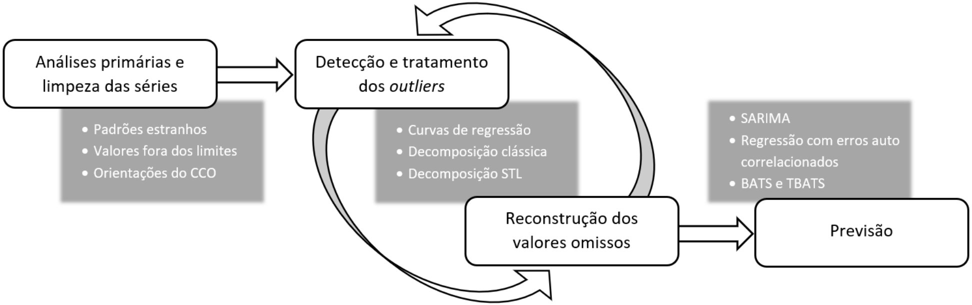
Leonardo Fonseca Larrubia[1], Chang Chiann[1] e Olga Satomi Yoshida[2]

[1]IME-USP - Instituto de Matemática e Estatística
[2]IPT - Instituto de Pesquisas Tecnológicas

Abril de 2021

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da FAPESP (Processo nº 2018/26592-5)

Esquema metodológico



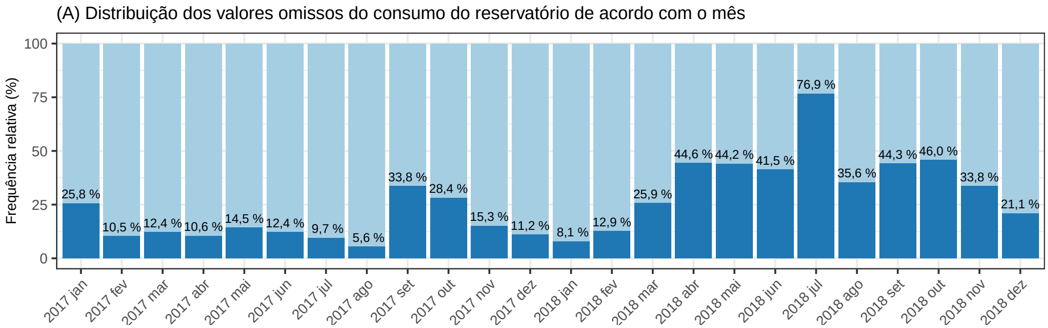
LIMPEZA DAS SÉRIES

Limpeza das séries

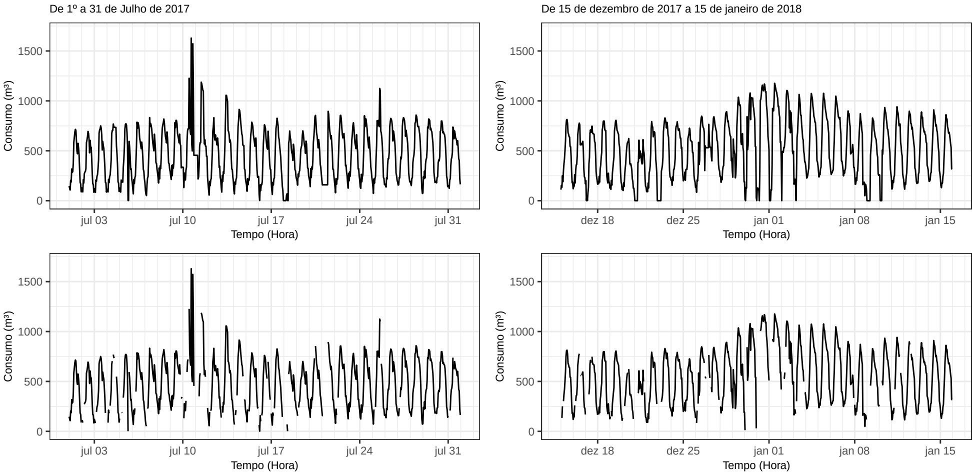
A limpeza primária das séries consiste em excluir:

- Valores fora dos limites máximo e mínimo da série de acordo com os parâmetros físico do sistema.
- Valores automaticamente interpolados pelo sistema de envio de dados.
- Valores considerados imprecisos de acordo com os técnicos do CCO da Sabesp.

Exemplo: Limpeza da série de consumo

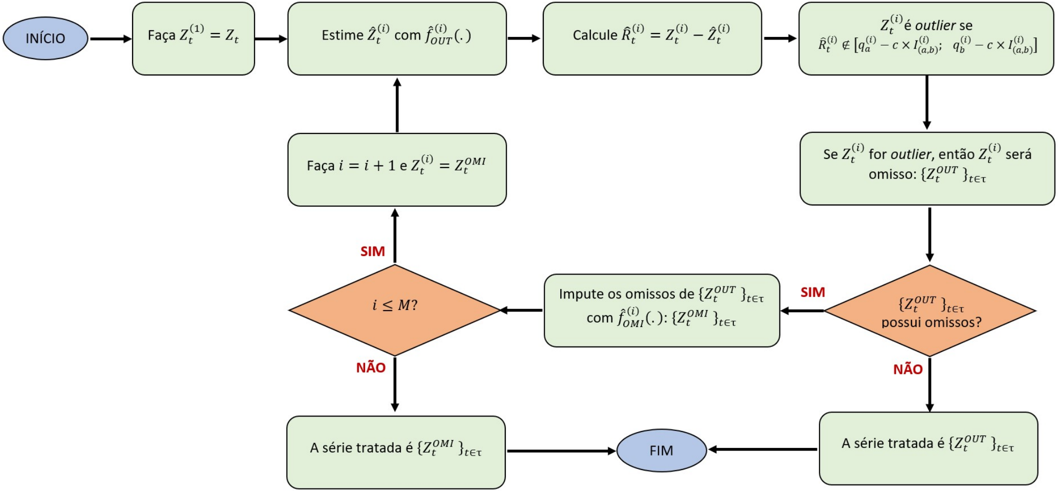


Exemplo: Limpeza da série de consumo



PROCESSO ITERATIVO DE TRATAMENTO DE OUTLIERS E VALORES OMISSOS

Processo iterativo de tratamento de *outliers* e valores omissos



Processo iterativo de tratamento de *outliers* e valores omissos

- Adotamos o mesmo procedimento $f_{OUT}^{(i)}(.) = f_{OUT}(.)$ para todos os $i = 1, ..., M$. A exceção é apenas em situações nas quais $f_{OUT}(.)$ não é robusto a valores omissos.
- Também adotamos o mesmo procedimento $f_{OMI}^{(i)}(.) = f_{OMI}(.)$ para todos os $i = 1, ..., M$.
- Tanto os métodos escolhidos para $f_{OUT}(.)$, quanto para $f_{OMI}(.)$ se baseiam em decompor a série nas componentes de tendência, sazonalidade e ruído.
    1. ajuste de curva de regressão;
    2. regressão combinado com decomposição clássica;
    3. decomposição STL.

AJUSTE DE CURVAS VIA REGRESSÃO

Procedimentos de ajuste de curvas via regressão

O modelo geral de regressão é:

$$Z_t = \sum_{i=0}^{m} \beta_i t^i + \sum_{j=1}^{23} \alpha_j D_{j,t} + \sum_{j=1}^{6} \gamma_j S_{j,t} + R_t, \tag{1}$$

em que

- m é o grau do polinômio usado para modelar a tendência;
- t^i é o i-ésimo termo do polinômio de grau m;
- β_i é o parâmetro associado ao i-ésimo grau do polinômio;
- α_j é o parâmetro associado ao j-ésimo horário do dia;
- $D_{j,t}$ é o j-ésimo horário associado ao tempo t;
- γ_j é o parâmetro associado ao j-ésimo dia da semana;
- $S_{j,t}$ é o j-ésimo dia da semana associado ao tempo t;
- R_t é o ruído do modelo.

Procedimentos de ajuste de curvas via regressão

O valor estimado para Z_t será

$$\hat{Z}_t^{Reg} = \sum_{i=0}^{m} \hat{\beta}_i t^i + \sum_{j=1}^{23} \hat{\alpha}_j D_{j,t} + \sum_{j=1}^{6} \hat{\gamma}_j S_{j,t} + \hat{R}_t, \quad (2)$$

em que os $\hat{\beta}_i$, os $\hat{\alpha}_j$ e os $\hat{\gamma}_j$ são as estimativas de mínimos quadrados. Em geral, $\hat{R}_t$ valerá 0, mas para os omissos também terá o valor dado pela equação:

$$\hat{R}_t = \hat{R}^{(0)} \left(1 - \frac{t - t^{(0)}}{t^{(1)} - t^{(0)}}\right) + \hat{R}^{(1)} \left(1 - \frac{t^{(1)} - t}{t^{(1)} - t^{(0)}}\right), \quad (3)$$

em que

- $\hat{R}^{(0)}$ é o último resíduos estimado antes do início da sequência de valores omissos;
- $\hat{R}^{(1)}$ é o primeiro resíduo estimado depois do término da sequência de valores omissos;
- $t^{(0)}$ é o índice de $\hat{R}^{(0)}$;
- $t^{(1)}$ é o índice de $\hat{R}^{(1)}$.

Procedimentos de ajuste de curvas via regressão

Utiliza-se subséries para os ajustes:

Detecção de *outliers*: As subséries são uma partição da série temporal de modo que cada subsérie tenha tamanho k, exceto a última subsérie, que poderá ter um tamanho entre $k/2$ e $3k/2$, porque não impomos que k seja um múltiplo da quantidade total de observações.

Imputação de omissos: Para cada sequência ininterrupta de valores omissos considerou-se a menor subsérie que contém os k valores não omissos mais próximos de tal sequência.

AJUSTE DE CURVAS VIA REGRESSÃO + DECOMPOSIÇÃO CLÁSSICA

Procedimentos de ajuste de curvas via regressão + decomposição clássica

Decomposição Clássica: Decompõe a série nas componentes de tendência, sazonalidade e ruído:

$$Z_t = \hat{T}_t + \hat{S}_t + \hat{R}_t. \tag{4}$$

- $\hat{T}_t$: é obtido por um filtro de média móveis de tamanho s (tamanho do período) à série $\{Z_t\}_{t \in \tau}$;
- $\hat{S}_t$: é obtido calculando a média de cada nível sazonal da série livre de tendência $(Z_t - \hat{T}_t)$;
- $\hat{R}_t$: é estimado por $\hat{R}_t = Z_t - \hat{T}_t - \hat{S}_t$.

Principais problemas:

- A sazonalidade estimada é a mesma durante toda a série, o que não seria uma premissa razoável para séries muito longas;
- A decomposição não consegue lidar com valores omissos.

Procedimentos de ajuste de curvas via regressão + decomposição clássica

A fase de **detecção de outliers**, $f_{OUT}^{(i)}(.)$, tem as seguintes característica:

- Na primeira iteração do algoritmo, $i = 1$, usa-se algum método de regressão para se fazer as estimativas $\hat{R}_t^{(1)}$ já explicado anteriormente;
- Nas demais iterações, usa-se o método de decomposição clássica para se calcular os resíduos $\hat{R}_t^{(i)}$, $i = 2, ..., M$. Para isso considerou-se uma partição da série temporal de modo que cada subsérie tenha tamanho k, exceto a última subsérie, que poderá ter um tamanho entre $k/2$ e $3k/2$, porque não impomos que k seja um múltiplo da quantidade total de observações;

Procedimentos de ajuste de curvas via regressão + decomposição clássica

Para a **imputação de valores omissos**, $f_{OMI}^{(i)}(.)$, para cada iteração i do algoritmo, tem-se duas etapas:

1ª Etapa: Ajuste de curvas e preenchimento dos valores omissos:

$$Z_t^{Reg} = \begin{cases} \hat{Z}_t^{Reg}, & \textit{se } Z_t \textit{ é omisso}; \\ Z_t, & \textit{caso contrário}; \end{cases}$$

2ª Etapa: Aplicação da decomposição clássica na série Z_t^{Reg}:

$$\hat{Z}_t^{Dcc} = \hat{S}_t + \hat{T}_t + \hat{R}_t,$$

no qual $\hat{S}_t$ e $\hat{T}_t$ são obtidos pelo procedimento de decomposição clássica à janela dos k valores mais próximos da sequência original de observações omissas da série Z_t^{Reg} e os $\hat{R}_t$ podem ou valer 0 ou ser o resultado de uma interpolação linear.

DECOMPOSIÇÃO STL

Procedimentos de decomposição STL

- STL, *Seasonal-Trend Decomposition Procedure Based on Loess*, proposto por Cleveland et al. (1990);
- Decompõe uma série temporal aditiva em três componentes: tendência, sazonalidade e ruído.

$$Z_t = \hat{T}_t + \hat{S}_t + \hat{R}_t, \tag{5}$$

- As **estimativas** são realizadas por um **processo iterativo** de dois *loops*:

  loop interno: Em cada iteração ambas a tendência e a sazonalidade são atualizadas através de aplicação sucessivas de suavizações *loess* em combinação com filtros de médias móveis.

  loop externo: Consiste de um *loop* interno seguido pelo cálculo de pesos robustos. No passo inicial todos os pesos são iguais a 1.

Procedimentos de decomposição STL

Para ambos os casos de detecção de *outliers* ou preenchimento de omissos, o valor estimado para Z_t é:

$$\hat{Z}_t^{Stl} = \hat{S}_t + \hat{T}_t + \hat{R}_t,$$

em que $\hat{S}_t$ e $\hat{T}_t$ são obtidos pela decomposição STL e para os $\hat{R}_t$ consideraremos os mesmo dois casos: $\hat{R}_t = 0$ e $\hat{R}_t$ como resultado da interpolação, sendo este só usado no preenchimento de valores omissos.

Discussão

- Os métodos propostos geraram resultados próximos entre si.
- A qualidade da imputação e detecção depende do comportamento da série na qual os métodos são aplicados.
- Seria interessante considerar o custo computacional de execução do procedimento.
- Os métodos propostos de imputação de valores omissos podem ser aperfeiçoados por uma mistura dos métodos de acordo com o tamanho da sequência de valores omissos.
- Possibilidade do uso dos métodos de previsão para imputação em sequências muito longas de omissos: *forescasting* + *backcasting*.
- Alta taxa de detecção de *outliers* pode significar uma detecção excessiva (falsa detecção);
- A escolha dos parâmetros de corte c e a quantidade máxima de iterações do algoritmo teriam que ser realizadas de modo mais subjetivo;

MODELOS DE PREVISÃO

Modelos de previsão

São propostos três tipos de modelos principais:

- SARIMA;
- Regressão com erros autocorrelacionados;
- BATS e TBATS.

MODELOS SARIMA

SARIMA: Modelo

Um modelo autorregressivo e de médias móveis sazonal multiplicativo, SARIMA $(p, d, q) \times (P, D, Q)_{[s]}$, é escrito da forma

$$\boldsymbol{\phi}(B)\,\boldsymbol{\Phi}(B^s)\,\Delta_s^D \Delta^d Z_t = \boldsymbol{\theta}(B)\,\boldsymbol{\Theta}(B^s)\,\xi_t, \qquad (6)$$

em que $\Delta_s^D \Delta^d Z_t$ é um processo estacionário; $E[Z_t] = 0$; $\xi_t \sim RB\left(0,\ \sigma^2\right)$ e

- $\boldsymbol{\phi}(B) = (1 - \phi_1 B - \ldots - \phi_p B^p)$ é o operador autorregressivo não sazonal;
- $\boldsymbol{\theta}(B) = (1 + \theta_1 B + \ldots + \theta_q B^q)$ é o operador de médias móveis não sazonal;
- $\boldsymbol{\Phi}(B^s) = \left(1 - \Phi_1 B^s - \ldots - \Phi_P B^{sP}\right)$ é o operador autorregressivo sazonal;
- $\boldsymbol{\Theta}(B^s) = \left(1 + \Theta_1 B^s + \ldots + \Theta_Q B^{sQ}\right)$ é o operador de médias móveis sazonal.

- **Estimação**: Pode ser realizada via máxima verossimilhança condicional usando a série $\{\Delta_s^D \Delta^d Z_t\}_{t \in \tau}$.
- **Previsão**: Usa-se o modelo estimado para fazer as previsões $\hat{Z}_{t+h|t}$.
- **Seleção de modelo**: algoritmo *stepwise* de Hyndman e Khandakar (2008).

MODELOS DE REGRESSÃO COM ERROS AUTOCORRELACIONADOS

Regressão com erros autocorrelacionados: Modelo

O modelo de regressão com erros autocorrelacionados é dado por

$$Z_t = \sum_{i=1}^{m} \beta_i x_{i,t} + W_t, \tag{7}$$

de modo que:

- W_t é um processo estacionário ARMA (p, q);
- $\sum_{i=1}^{m} \beta_i x_{i,t}$ é construído de forma a
  - considerar sazonalidade horária;
  - pode considerar ou não sazonalidade semanal;
  - pode considerar tendência constante ou linear.

- **Estimação**: Pode ser realizada mínimos quadrados.
- **Previsão**: Usa-se o modelo estimado para fazer as previsões $\hat{Z}_{t+h|t}$.
- **Seleção de modelo**: Algoritmo *stepwise* de Hyndman e Khandakar (2008).

MODELOS BATS e TBATS

BATS e TBATS

- Propostos por De Livera et al. (2011);
- BATS: Box-Cox transformation, ARMA erros, Trend and Seasonal components;
- TBATS: Trigonometric, Box-Cox transformation, ARMA erros, Trend and Seasonal components;
- São extensões dos modelos de suavização exponencial;
- Buscam lidar com séries temporais que apresentem "comportamento complexo":
  - sazonalidade cuja frequência é um número não inteiro;
  - séries com frequências muito longas;
  - presença de mais de uma sazonalidade.

BATS e TBATS: Transformação Box-Cox

A transformação Box-Cox é:

$$Z_t^{(\omega)} = \begin{cases} \dfrac{(Z_t)^{\omega} - 1}{\omega}, & \omega \neq 0; \\ \log Z_t, & \omega = 0. \end{cases} \tag{8}$$

- Existem diversas formas para a escolha do melhor parâmetro ω.
- O cálculo de ω é realizado de acordo com Guerreiro (1993): ω que minimiza o coeficiente de variação das subséries de $\{Z_t\}_{t \in \tau}$.

BATS: Modelo

O modelo BATS $(\omega,\ \varphi,\ p,\ q,\ m_1, \ldots,\ m_T)$, com T padrões sazonais, é

$$Z_t^{(\omega)} = \ell_{t-1} + \varphi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t, \tag{9}$$

em que:

$$\begin{aligned}
\ell_t &= \ell_{t-1} + \varphi b_{t-1} + \alpha d_t; \\
b_t &= (1-\varphi)\, b + \varphi b_{t-1} + \beta d_t; \\
s_t^{(i)} &= s_{t-m_i}^{(i)} + \gamma_i d_t; \\
d_t &= \textstyle\sum_{i=1}^{p} \phi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \xi_{t-i} + \xi_t;
\end{aligned} \tag{10}$$

TBATS: Modelo

O modelo TBATS $(\omega,\ \varphi,\ p,\ q,\ \{m_1, k_1\}, \ldots,\ \{m_T, k_T\})$, com T padrões sazonais, é

$$Z_t^{(\omega)} = \ell_{t-1} + \varphi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t, \tag{11}$$

em que:

$$\begin{aligned}
\ell_t &= \ell_{t-1} + \varphi b_{t-1} + \alpha d_t; \\
b_t &= (1-\varphi)\, b + \varphi b_{t-1} + \beta d_t; \\
s_t^{(i)} &= \textstyle\sum_{j=1}^{k_i} s_{j,t}^{(i)}; \\
s_{j,t}^{(i)} &= s_{j,t-1}^{(i)} \cos \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin \lambda_j^{(i)} + \gamma_1^{(i)} d_t; \\
s_{j,t}^{*(i)} &= -s_{j,t-1} \sin \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos \lambda_j^{(i)} + \gamma_2^{(i)} d_t; \\
d_t &= \textstyle\sum_{i=1}^{p} \phi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \xi_{t-i} + \xi_t;
\end{aligned} \tag{12}$$

BATS e TBATS: Estimação seleção de modelos e Previsão

- **Estimação**: Máxima verossimilhança condicionada ou os estimadores que minimizam a soma dos erros ao quadrado (modelos na forma de equações de estado-espaço);
- **Previsões**: Também são derivadas das equações de estado-espaço através do filtro de Kalman;
- **Seleção de modelo**: Os autores propõem um procedimento que usa o AIC;
- **Seleção do modelo ARMA para o ruído**: Procedimento em três estágios:
  1. Estima-se o modelo supondo ruído branco;
  2. Seleciona-se um modelo ARMA aos ruídos de acordo com o algoritmo *stepwise*;
  3. Com o modelo ARMA selecionado, todo o modelo BATS ou TBATS é reajustado considerando o processo ARMA selecionado.

Discussão

- A qualidade da previsão depende muito do comportamento da série na qual os métodos são aplicados.
- Seria interessante considerar o custo computacional.
- O uso em tempo real, além de auxiliar o operador a tomar decisões sobre a demanda de água, também pode servir para detecção de anomalias geradas por algum erro no sistema ou por um consumo atípico da população.