

Modelo PMIX

Leandro Furlam Turi

Mônica de Souza Mendes Castro

22 de Maio de 2019

1 Introdução

Conhecer o comportamento de séries hidrológicas é de fundamental importância para diferentes finalidades no planejamento e gerenciamento de recursos hídricos. Geralmente, séries históricas de vazões são sintéticas curtas, o que torna necessária a utilização de séries sintéticas que contenham as mesmas propriedades estatísticas da série original, porém simulando sequências maiores de possíveis realizações de eventos de cheias e secas. Para resolver essas lacunas, (MENDONÇA, 1987) desenvolveu o modelo *multiplicative Periodic AutoRegressive Moving Average*, denominado PMIX(p,q,P,Q).

Definimos o modelo da seguinte forma:

$x_{v,\tau}$ o valor de vazão medido de um determinado mês τ em um determinado ano v , com média μ_τ e desvio-padrão σ_τ

$$y_{v,\tau} = \frac{\ln(x_{v,\tau}) - \mu_\tau}{\sigma_\tau}$$

$$\begin{aligned} y_{v,\tau} &= \phi_{1,\tau} y_{v,\tau-1} + \phi_{2,\tau} y_{v,\tau-2} + \dots + \phi_{p,\tau} y_{v,\tau-p} \\ &+ \Phi_{1,\tau} y_{v-1,\tau} + \Phi_{2,\tau} y_{v-2,\tau} + \dots + \Phi_{P,\tau} y_{v-P,\tau} \\ &- \phi_{1,\tau} \Phi_{1,\tau} y_{v-1,\tau-1} - \phi_{1,\tau} \Phi_{2,\tau} y_{v-2,\tau-1} - \dots \\ &- \phi_{1,\tau} \Phi_{P,\tau} y_{v-P,\tau-1} - \dots - \phi_{p,\tau} \Phi_{1,\tau} y_{v-1,\tau-p} \\ &- \phi_{p,\tau} \Phi_{2,\tau} y_{v-2,\tau-p} - \dots - \phi_{p,\tau} \Phi_{P,\tau} y_{v-P,\tau-p} \\ &+ \varepsilon_{v,\tau} \\ &= \theta_{1,\tau} \varepsilon_{v,\tau-1} + \theta_{2,\tau} \varepsilon_{v,\tau-2} + \dots + \theta_{q,\tau} \varepsilon_{v,\tau-q} \\ &+ \Theta_{1,\tau} \varepsilon_{v-1,\tau} + \Theta_{2,\tau} \varepsilon_{v-2,\tau} + \dots + \Theta_{Q,\tau} \varepsilon_{v-Q,\tau} \\ &- \theta_{1,\tau} \Theta_{1,\tau} \varepsilon_{v-1,\tau-1} - \theta_{1,\tau} \Theta_{2,\tau} \varepsilon_{v-2,\tau-1} - \dots \\ &- \theta_{1,\tau} \Theta_{Q,\tau} \varepsilon_{v-Q,\tau-1} - \dots - \theta_{q,\tau} \Theta_{1,\tau} \varepsilon_{v-1,\tau-q} \\ &- \theta_{q,\tau} \Theta_{2,\tau} \varepsilon_{v-2,\tau-q} - \dots - \theta_{q,\tau} \Theta_{Q,\tau} \varepsilon_{v-Q,\tau-q} \end{aligned}$$

Onde os parâmetros $\phi_{1,\tau}, \phi_{2,\tau}, \dots, \phi_{p,\tau}$ são os coeficientes autorregressivos periódicos de *lag 1*, *lag 2*, ..., *lag p*, respectivamente, e $\Phi_{1,\tau}, \Phi_{2,\tau}, \dots, \Phi_{P,\tau}$ são os coeficientes autorregressivos periódicos de *lag 1w*, *lag 2w*, ..., *lag Pw*.

Os parâmetros $\theta_{1,\tau}, \theta_{2,\tau}, \dots, \theta_{p,\tau}$ são os coeficientes médias móveis de *lag 1*, *lag 2*, ..., *lag q*, respectivamente, e $\Theta_{1,\tau}, \Theta_{2,\tau}, \dots, \Theta_{P,\tau}$ são os coeficientes médias móveis de *lag 1w*, *lag 2w*, ..., *lag Qw*.

$y_{v,\tau}$ e $\varepsilon_{v,\tau}$ apresentam média zero e variâncias $\sigma_\tau^2(y)$ e $\sigma_\tau^2(\varepsilon)$, respectivamente.

2 Algoritmo

Na busca da estimação ótima dos parâmetros ϕ, θ, Φ e Θ no modelo, foi implementada a heurística evolutiva Algoritmo Genético, com sua forma proposta pela técnica *fast Nondominated Sorting Genetic Algorithm*

(NSGA-II). O algoritmo foi escrito na linguagem R¹, com o auxílio de pacotes-padrão disponíveis no software.

A saber, Algoritmos genéticos diferem dos algoritmos tradicionais de otimização em basicamente quatro aspectos (GOLDBERG; HOLLAND, 1988):

- Baseiam-se em uma codificação do conjunto das soluções possíveis, e não nos parâmetros da otimização em si;
- Os resultados são apresentados como uma população de soluções e não como uma solução única;
- Não necessitam de nenhum conhecimento derivado do problema, apenas de uma forma de avaliação do resultado;
- Usam transições probabilísticas e não regras determinísticas.

Como forma de restringir o espaço de busca e melhorar o tempo de processamento do algoritmo, foi utilizado como valores iniciais resultados obtidos pelo método de Powell, onde seus pontos de partida foram gerados de forma aleatória, buscando varrer o máximo do espaço de busca.

A seguir são descritos os métodos e métricas utilizadas na implementação.

2.1 Funções de avaliação

2.1.1 Média

$$MAPE_{media} = \frac{\sum_{\omega=1}^{12} \left| \frac{\mu_{\omega}^{hist} - \mu_{\omega}^{sint}}{\mu_{\omega}^{hist}} \right|}{12}$$

2.1.2 Desvio-padrão

$$MAPE_{dp} = \frac{\sum_{\omega=1}^{12} \left| \frac{\sigma_{\omega}^{hist} - \sigma_{\omega}^{sint}}{\sigma_{\omega}^{hist}} \right|}{12}$$

2.1.3 Função de autocorrelação anual

$$MAPE_{facAnual} = \frac{\sum_{a=1}^{max(lag_a)} \left| \frac{\rho_a^{hist} - \rho_a^{sint}}{\rho_a^{hist}} \right|}{lag_a^{max}}$$

2.1.4 Função de autocorrelação mensal

$$MAPE_{facMensal} = \frac{\sum_{m=1}^{12} \sum_{i=1}^{max(lag_m)} \left| \frac{\rho_{m,i}^{hist} - \rho_{m,i}^{sint}}{\rho_{m,i}^{hist}} \right|}{12 * max(lag_m)}$$

2.1.5 Somatório dos resíduos

$$SomRes = \sum_{\tau=1}^{10000} \sum_{\omega=1}^{12} \epsilon^2$$

2.2 Medidas genéticas sugeridas

Nesta seção são apresentados as medidas que melhor se adaptaram e resultaram em melhores resultados na maioria dos testes realizados.

¹Software livre, disponível em <https://www.R-project.org/>

2.2.1 Parâmetros Genéticos

- Tamanho da população: 50 indivíduos;
- Probabilidade de cruzamento: 80%;
- Probabilidade de mutação: 20%.

2.2.2 Critérios de parada

- Número máximo de ciclos: 10 mil;
- MAPE máximo da diferença entre parâmetros: 20%. Este critério de parada relaciona-se à convergência das soluções.

3 Funcionamento

A implementação possui duas formas de estimar os parâmetros: uma com ordens isoladas e outra completa, com todos as ordens que normalmente são encontrados na literatura. A primeira é mais rápida, e é indicada quando se conhece as ordens do modelo que melhor se aplicam a estação analisada. A segunda é indicada quando se deseja fazer uma análise completa da estação.

A seguir são apresentadas tais funções e o tipo de cada variável de entrada.

3.1 Dicionário

- p: ordem autorregressiva mensal;
- q: ordem media movel mensal;
- P: ordem autorregressiva anual;
- Q: ordem media movel anual;
- lags: (p, q, P, Q);
- dados: caminho do arquivo;
- n: tamanho da serie sintética a ser gerada.
- nP: tamanho da população;
- Pc: probabilidade de cruzamento;
- Pm: probabilidade de mutação;
- cicloMAX: ciclo Maximo do AG;
- MAXDiferenca: diferença máxima entre os parâmetros encontrados pelo AG;

3.2 Geração dos parâmetros

3.2.1 Ordem específica (arquivo tempo.R)

Durante a interpretação dos códigos, será solicitado que escolha o arquivo contendo a série histórica.

```
# lags: numeric (4)
# nP: numeric (1)
# Pc: numeric (1)
# Pm: numeric (1)
# cicloMAX: numeric (1)
# MAXDiferenca: numeric (1)

NSGA (lags , nP , Pc , Pm , cicloMAX , MAXDiferenca)
```

3.2.2 Análise completa (arquivo modeloPMIX.R)

```
# dados: character (Caminho do arquivo)
#
# lags (p,q,P,Q) utilizados
# (1,0,0,0)
# (1,0,1,0)
# (1,1,0,0)
# (1,1,1,0)
# (1,1,1,1)
# (2,0,0,0)
# (2,0,1,0)
# (2,1,0,0)
# (2,1,1,0)
# (2,1,1,1)
```

```
modeloPMIX (dados)
```

3.3 Medidas estatísticas

3.3.1 Funções de autocorrelação (arquivo correlograma.R)

Como saídas, será obtida a matriz de autocorrelações, onde as linhas correspondem aos lags, e as colunas correspondem aos meses, iniciando do lag 0.

```
# serie: matrix (ncol = nMeses, nrow = nAnos)
# lagMax = numeric (1)
```

```
autocorrelacaoMensal (serie, lagMax)
```

Como saída, será obtida um vetor contendo as autocorrelações anuais, iniciando do lag 0.

```
# serieAnual: numeric (nAnos)
# lagMax = numeric (1)
```

```
autocorrelacaoAnual (serieAnual, lagMax)
```

3.3.2 Geração do cenário sintético (arquivo CenarioSintetico.R)

Seguindo (SALAS; DELLEUR; YEVJEVICH, 1980), são descartados os primeiros 50 anos da série sintética, referentes ao aquecimento do modelo.

Como saída, será obtida a série sintética gerada pelos parâmetros informados.

```
# serieH = matrix (ncol = nMeses, nrow = nAnos)
# parametros = numeric (nMeses*(p+q+P+Q))
# lags: numeric (4)
# n = numeric (1)
```

```
cenarioSintetico (serieH, parametros, lags, n)
```

3.3.3 Cálculo do somatório dos resíduos (arquivo sumQuadRes.R)

Como saída, será indormada a soma dos quadrados dos resíduos e o desvio padrão residual.

```
# serie: matrix (ncol = nMeses, nrow = nAnos)
# parametros = numeric (nMeses*(p+q+P+Q))
# lags: numeric (4)
```

```
residuos (serie, parametros, lags)
```

3.3.4 Cálculo dos parâmetros através do método de Powell (arquivo powell.R)

Durante a interpretação dos códigos, será solicitado que escolha o arquivo contendo a série histórica.

```
# p: numeric (1)
# q: numeric (1)
# P: numeric (1)
# Q: numeric (1)

PMIX (p, q, P, Q)
```

3.3.5 Geração de cenários anuais (arquivo CenarioAnual.R)

Como saída, serão informados os parâmetros encontrados pelo método de redução do somatório dos quadrados dos resíduos, os resíduos gerados, e a série sintética gerada.

```
# P: numeric (1)
# Q: numeric (1)

cenarioSinteticoAnual (P, Q)
```

4 Clusterização

Devido a quantidade de operações e dados gerados, foi optado por utilizar técnicas de paralelismo. Para isso, é feito o uso do pacote **Parallel**².

O próprio algoritmo já resolve todas as dependências, sendo necessário ao pesquisador apenas definir a quantidade de nós para a técnica de *clusterização* feita pelo pacote, através da variável `cores`.

ATENÇÃO: devido a restrições de software, usuários de **Windows devem utilizar apenas 1 nó** para que as funções do pacote funcionem corretamente. Os autores deste modelo recomendam a utilização do sistema operacional Linux para essa aplicação.

```
require ('parallel')
cores = 1
cl = makeCluster (cores)
```

5 Saídas

Após o término da aplicação, o algoritmo terá como saída uma pasta localizada no diretório onde se encontrava o arquivo contendo a série histórica, com o nome da estação. Nessa pasta encontram-se subpastas com saídas referentes a todas as ordens de modelo analisadas. Nessas, encontram-se os seguintes arquivos e subpastas:

- TempoExecucao.txt: arquivo com o tempo de execução e o número de ciclos necessários para aquela ordem de modelo;
- ParametrosIniciais.csv: arquivo com os parâmetros iniciais obtidos pelo método de Powell utilizados como ponto inicial pelo Algoritmo Genético;
- ParametrosFinais.csv: arquivo com os parâmetros finais encontrados;
- Series: subpasta contendo as séries sintéticas geradas pelo conjunto de parâmetros encontrados como solução;
- Avaliaco.es.csv: arquivo com as avaliações dos indivíduos encontrados como ótimos.

²<https://www.rdocumentation.org/packages/parallel>

Referências

GOLDBERG, D. E.; HOLLAND, J. H. *Genetic algorithms and machine learning*. [S.l.]: Springer, 1988. v. 3. 95–99 p.

MENDONÇA, A. S. F. *Stochastic modeling of seasonal streamflow*. 246 p. Tese (Doutorado) — Colorado State University, 1987.

SALAS, J.; DELLEUR, J. W.; YEVJEVICH, V. *Applied modeling of hydrologic time series*. Colorado: Book Crafters, 1980.