

Comparação de modelos de previsão na série temporal de instrumento de monitoramento de barragem

Emerson Lazzarotto¹

Vitoria Eduarda Bello²

Kelly Maiara Masur da Silva³

UNIOESTE, Univ. Est. Oeste do Paraná, Foz do Iguaçu, PR

Resumo. Dada a série temporal que monitora alguma variável de interesse em uma barragem é importante saber se a característica que está sendo avaliada continua tendo comportamento normal no período operacional da barragem. Neste trabalho compara-se duas técnicas probabilísticas de modelagem baseadas nos dados da própria série, modelo SARIMA e de regressão linear múltipla, com o objetivo de avaliar a eficiência da previsão da série de um instrumento que monitora uma barragem. Os resultados mostraram que, em geral, os modelos SARIMA tiveram menor erro de previsão (6% em média) e podem colaborar com a equipe de engenharia como uma ferramenta adicional à segurança de barragens.

Palavras-chave. Modelos SARIMA, Modelos de regressão, previsão de séries temporais, segurança de barragem

1 Introdução

As barragens possuem diversas formas, tamanhos e finalidades. Uma das mais importantes finalidades são as barragens de usina de geração hidrelétricas [3]. No período operacional de uma usina é fundamental continuar a monitorar o comportamento da barragem para assegurar sua segurança [9]. Seja por conta de acidentes ou não, falhas na segurança, como infelizmente já vistos no Brasil, podem causar grande impacto econômico, ambiental e social.

Quando se fala em monitoramento de uma barragem no seu período operacional este se dá basicamente por conta de inspeções visuais e instrumentos de monitoramento [8]. Os instrumentos são instalados em locais específicos que se deseja avaliar o comportamento da barragem com o passar do tempo.

A função destes instrumentos é medir características (variáveis) de engenharia importantes como vazão, subpressão, deslocamento e recalque entre 'blocos' da barragem. As leituras periódicas de um instrumento geram séries temporais de medidas de uma variável.

Mesmo funcionando em perfeita normalidade, um aspecto que os dados de uma série temporal de monitoramento de barragem pode fornecer são eventuais tendências através da realização de algum método que execute previsões. Qualquer tipo de indicação prévia que ajude a antecipar ações que mitiguem riscos e custos é extremamente útil à equipe de segurança de barragem.

Devido à complexidade e ao tamanho de uma 'grande' barragem, modelos determinísticos de previsão podem não ser mais adequados na fase operacional da barragem [2]. A existência de

¹emerson.lazzarotto@gmail.com

²vitoriaeduardabello@hotmail.com

³Kelly.masur@hotmail.com

dados de monitoramento tem possibilitado a aplicação de modelos estatísticos para a previsão do comportamento [7].

Portanto, neste trabalho realiza-se a previsão da série temporal de um instrumento de monitoramento da barragem da usina hidrelétrica de Itaipu por meio de dois modelos estatísticos paramétricos clássicos: modelos sazonais autorregressivos integrados de média móvel (SARIMA) e modelos de regressão linear múltipla. Busca-se fazer uma comparação entre diversos modelos e a determinação de um modelo 'ótimo' no sentido de produzir as previsões mais acuradas.

2 Referencial teórico

2.1 Modelos SARIMA

Dada uma série temporal z_1, z_2, \dots, z_N , de N observações com intervalos de tempo igualmente espaçados, o operador de retardo B é definido por $Bz_t = z_{t-1}$, de modo que, $B^m z_t = z_{t-m}$. Para séries temporais com sazonalidade cujo período é S utiliza-se o operador $B^S z_t = z_{t-S}$. Além disso, o operador de diferença, ∇ , é definido por $\nabla z_t = z_t - z_{t-1} = (1 - B)z_t$ (diferença de ordem 1) e ∇_S , é definido por $\nabla_S z_t = z_t - z_{t-S} = (1 - B^S)z_t$ (diferença sazonal de ordem S).

Segundo [1], os modelos sazonais multiplicativos autorregressivos integrados de médias móveis (SARIMA) têm a forma geral

$$\phi_p(B)\Phi_P(B^S)\nabla^d\nabla_S^D z_t = \theta_q(B)\Theta_Q(B^S)a_t. \quad (1)$$

Na equação (1), S é a sazonalidade (12, por exemplo, para dados mensais), $\Phi_P(B^S)$ e $\Theta_Q(B^S)$ são polinômios em B^S de graus P e Q , respectivamente; $\phi_p(B)$ e $\theta_q(B)$ são polinômios em B de graus p e q , respectivamente; $\nabla_S = 1 - B^S$ e $\nabla = 1 - B$; D e d são as respectivas ordens de diferenciação sazonal e diferença simples e os a_t é uma série de 'choques' aleatórios ou ruído branco [1]. O modelo da equação (1) é dito ser de ordem $(p, d, q) \times (P, D, Q)_S$.

A obtenção de um modelo para uma série temporal segue uma estratégia baseada num processo iterativo no qual a escolha da estrutura do modelo se baseia nos próprios dados. As fases do processo consistem na especificação de classe geral de modelos, identificação de um modelo com base na análise de autocorrelações e autocorrelações parciais, estimação dos parâmetros do modelo e verificação, através da análise dos resíduos do modelo, no sentido de fornecer o menor erro de previsão [6].

2.2 Modelos de regressão linear múltipla

Em geral, um modelo de regressão linear múltipla com k variáveis independentes tem a forma

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon. \quad (2)$$

Na equação (2), y é a variável dependente ou regredida e as x_j são as variáveis regressoras ou independentes, β_j , $j = 0, 1, \dots, k$, são os parâmetros do modelo e ϵ é o termo do erro do modelo que, segundo [5], devem ser normais com valor esperado 0, variância constante e os ϵ_i , $i = 1, 2, \dots, n$, onde n é o tamanho da amostra, são não autocorrelacionados, em resumo, $\epsilon_i \sim NID(0, \sigma^2)$.

A partir de um conjunto de dados das variáveis independentes e da variável dependente pode-se determinar uma estimativa para os parâmetros β_j , como, por exemplo, através do método de mínimos quadrados.

A bem conhecida estatística R^2 , que varia de 0 a 1, é uma medida da quantidade de redução na variabilidade de y obtida pelo uso das variáveis de regressão x_1, x_2, \dots, x_k no modelo. Embora, um valor alto de R^2 não quer dizer que, necessariamente, o modelo seja 'bom'. Não obstante, pode ser usada para medir a 'qualidade' do ajustamento.

Para a análise da significância estatística da regressão globalmente e de coeficientes específicos de cada variável independente podem ser aplicados: testes de hipóteses sobre o conjunto de dados com o objetivo de determinar a utilidade de cada variável independente sobre o modelo, determinar a média esperada da variável dependente para cada n -upla de variáveis independentes e realizar previsões [5].

A verificação da adequação de um modelo de regressão linear múltipla é parte fundamental do procedimento da análise de dados. De acordo com [5], devem ser examinados os gráficos de resíduos em um modelo de regressão. O exame do modelo ajustado é necessário para se garantir que ele fornece uma aproximação adequada para o verdadeiro sistema e a verificação de que nenhuma das hipóteses de regressão foi violada.

Para a análise de resíduos e seus gráficos costuma-se levar em consideração os denominados Resíduos '*Studentizados*'. Em geral, os resíduos '*Studentizados*' são escalonados e, desta forma, transmitem mais informação do que os resíduos ordinários. Uma vez que o escalonamento faz com que sejam padronizados: com média 0 e variância 1, a 'maioria' dos resíduos '*Studentizados*' fica no intervalo entre -3 e 3 e, assim, eles são úteis na busca de valores atípicos. Segundo [5], quando existirem, devem ser estudados e investigada a possibilidade de excluir tais pontos da análise. Pode-se encontrar a fórmula do seu cálculo em [5].

3 Materiais e métodos

A aplicação deste trabalho deu-se com a série temporal de leituras de um instrumento de monitoramento da barragem da usina hidrelétrica de Itaipu Binacional, a maior geradora de energia limpa e renovável do mundo. Mais especificamente, este instrumento chama-se base de alongâmetro. Sua função é medir os deslocamentos diferenciais entre os blocos da barragem principal, nas juntas de contração transversais [4]. Estes blocos, em geral, são estruturas de concreto com dimensões de dezenas de metros.

As bases de alongâmetro estão instaladas em dois planos, um vertical e outro horizontal e possibilitam 3 tipos de controle de deslocamentos entre blocos: a *abertura* e o fechamento das juntas, o *deslizamento* na direção horizontal e o *recalque* diferencial entre blocos. Neste trabalho, escolheu-se uma série temporal que mede a abertura entre dois blocos. Esta abertura avalia a abertura das juntas provocada pela retração térmica e o fechamento provocado por deformações [4].

Para investigar o possível efeito da temperatura ambiente sobre esta variável de engenharia, a *abertura*, resolveu-se usar um modelo de regressão linear múltipla. Havia a disponibilidade de quatro termômetros, denominados de T_1 a T_4 . Os dados destas cinco séries temporais consistiam de uma média mensal do período de janeiro/1984 a maio/2016, ou seja, 389 dados para cada série.

Como um dos critérios de escolha do 'melhor' modelo separou-se uma parte final da amostra sempre com 12 dados (período de um ano) para validação dos modelos determinados na parte inicial da amostra, o período de estimação. Os modelos foram obtidos no *software Statgraphics*. Por exemplo, na Tabela 1 exemplifica-se os respectivos períodos de estimação e validação para a série completa de dados.

Tabela 1: Separação da amostra completa.

Etapa	Período	Número de dados
Estimativa	Jan/1984 a Mai/2015	377
Validação	Jun/2015 a Mai/2016	12

Com o intuito de determinar um período em que as medidas de desempenho do erro de previsão (no período de estimativa e de validação) tivessem o melhor resultado possível, fixou-se os doze

dados finais da amostra para validação e foram gerados diversos cenários conforme a Tabela 2 para o período de estimativa. Por exemplo, o cenário 1 com 377 dados corresponde à série completa. A razão para isto é que a série temporal pode ter tido um comportamento inicial que não mais reflete seu comportamento cíclico atual. A medida de erro usada no processo de decisão foi o erro absoluto médio percentual (MAPE), a ser usado daqui por diante.

Tabela 2: Cenários fixados no período de estimativa.

Cenário	1	2	3	4	5	6	7	8	9
N° dados	377	352	327	302	277	252	227	202	177

4 Resultados

4.1 Resultados da modelagem SARIMA

A Tabela 3 apresenta os melhores modelos encontrados, segundo o *software* utilizado, em cada cenário. A Tabela 4 apresenta os resultados do MAPE(%) nos períodos de estimativa e validação e se houve adequação (sim/não) do melhor modelo determinado em cada cenário em relação a todas as hipóteses teóricas subjacentes dos modelos, dentre elas: inexistência de autocorrelações e autocorrelações parciais até 12 *lags* de retardo, teste de aleatoriedade e normalidade dos resíduos e que os coeficientes do modelo fossem todos estatisticamente significativos em relação à determinação do melhor cenário de previsão para os modelos SARIMA.

Tabela 3: Modelo SARIMA determinado em cada cenário.

Cenário	Modelo SARIMA
1 a 3	$(2, 0, 2) \times (1, 1, 2)_{12}$
4	$(2, 1, 1) \times (2, 1, 2)_{12}$
5 e 6	$(1, 1, 1) \times (2, 1, 2)_{12}$
7	$(1, 0, 0) \times (1, 1, 2)_{12}$
8	$(1, 1, 2) \times (1, 1, 2)_{12}$
9	$(1, 0, 0) \times (2, 1, 2)_{12}$

Tabela 4: MAPE (%) e adequação para os melhores modelos SARIMA por cenário.

Cenário	1	2	3	4	5	6	7	8	9
MAPE _{est}	ND	8,14	7,83	7,71	7,13	6,99	7,06	6,82	6,71
MAPE _{val}	ND	6,60	6,97	7,20	7,01	6,97	6,61	6,42	6,67
Adequado	sim	não	não	não	sim	não	sim	não	sim

Em geral, a não adequabilidade de alguns modelos 'ótimos' em alguns cenários esteve relacionada com a existência de algum coeficiente estatisticamente não significativo no modelo. O termo modelo 'ótimo' é no sentido de que foram obtidos pela execução de comando automático no *software* sem a intervenção dos autores. A intervenção manual de inserir ou retirar termos do modelo é possível, entretanto, torna a busca muito mais trabalhosa, além disso, alguns testes realizados mostraram que não há garantias que gerariam um modelo adequado com melhor MAPE. No caso

do cenário 1, não foi possível calcular o MAPE devido à presença de dados com valores negativos no início da série.

4.2 Resultados da regressão múltipla

O efeito das temperaturas medidas sobre os termômetros T_i foi avaliado não só no instante de tempo t , mas, também, nos instantes de tempo $t - 1$, $t - 2$ e $t - 3$, ou seja, a influência da temperatura de até 3 meses antes sobre a leitura da série da base de alongâmetro. Estas variáveis serão indicadas por $T_i(-j)$, onde $i = 1, \dots, 4$ e $j = 1, \dots, 3$.

Assim, foram levadas em conta até 16 variáveis em modelos com, no máximo, 6 variáveis regressoras mais a constante (por causa do tempo computacional), totalizando 14.893 modelos ajustados para os diversos conjuntos de dados fixados em cada cenário já descrito.

Por exemplo, para o cenário 1, chegou-se à conclusão que o modelo que apresenta o maior grau de ajustamento R^2 é o modelo que leva em consideração as variáveis T_2 , T_3 , $T_2(-1)$, $T_2(-2)$, $T_3(-2)$ e $T_3(-3)$.

Para a regressão linear múltipla envolvendo os cenários em que a série temporal da base de alongâmetro está em função das séries temporais de 4 termômetros que estão localizados próximos, a Tabela 5 apresenta os resultados do grau de ajustamento R^2 no período de estimativa e o MAPE no período de validação e, também, a adequabilidade às hipóteses subjacentes de modelos de regressão múltipla citados na seção 2.2.

Tabela 5: R^2 e MAPE dos melhores modelos de regressão linear múltipla encontrados por cenário.

Cenário	1	2	3	4	5	6	7	8	9
R^2 (%)	81,27	86,51	88,79	89,96	91,81	92,36	93,72	94,16	94,29
MAPE (%)	24,21	21,80	21,98	22,32	21,11	21,14	19,68	19,46	18,57
Adequado	sim	sim	sim	sim	sim	sim	sim	não	não

Nos modelos de regressão múltipla, também, a não adequabilidade nos cenários 8 e 9, esteve relacionada com a existência de algum coeficiente estatisticamente não significativo para o modelo.

4.3 Comparação das previsões

Conclui-se, portanto, que dentre todos os modelos *teoricamente adequados*, segundo os cenários estabelecidos, apresentados nas Tabelas 4 e 5, aquele que apresentou o melhor desempenho de previsão com um horizonte de 12 meses, através da medida do MAPE foi o modelo SARIMA $(1, 0, 0) \times (1, 1, 2)_{12}$, observado no cenário 7.

Na Figura 1 exibem-se os dados reais e a previsão obtida segundo o melhor modelo determinado, ou seja, o modelo SARIMA $(1, 0, 0) \times (1, 1, 2)_{12}$. Na Figura 2 exhibe-se um gráfico comparativo do erro relativo % em cada mês do período preditivo usando o melhor modelo SARIMA e o melhor modelo de regressão linear múltipla encontrados.

5 Conclusões

Dada uma série temporal de dados que monitoram o deslocamento entre blocos em um trecho da barragem da usina hidrelétrica de Itaipu, o objetivo deste trabalho foi determinar um modelo probabilístico e um período da série que melhor executa-se previsões.

Dentre os tipos de modelos SARIMA e de regressão linear múltipla e os períodos de tempo selecionados através de 9 cenários estipulados, o modelo que apresentou os menores valores de

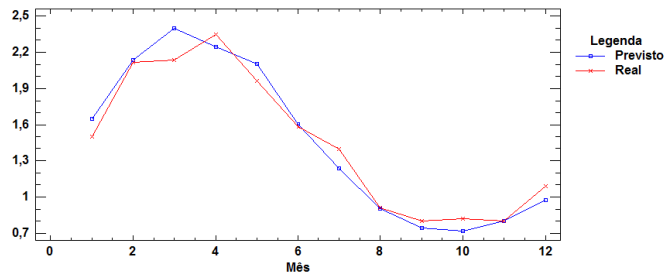


Figura 1: Comparação das leituras reais e previstas (cenário 7 modelo SARIMA).

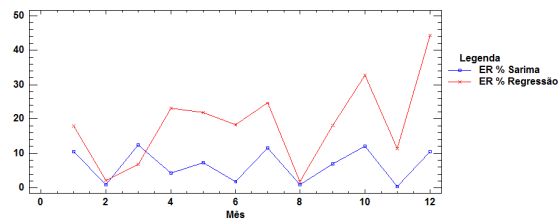


Figura 2: Comparação do erros relativos percentuais de previsão dos melhores modelos SARIMA x regressão.

erro (MAPE) satisfazendo as hipóteses básicas de adequação foi um SARIMA $(1, 0, 0) \times (1, 1, 2)_{12}$ consistindo dos últimos 239 dados (sendo 12 retidos para o período de validação).

Embora tenha havido gradual melhora do grau de ajustamento com a utilização de dados mais atuais (últimos cenários) os modelos de regressão linear múltipla apresentaram desempenho de previsão bem inferior aos modelos SARIMA. Uma razão para isto pode ser o registro de temperaturas atípicas em determinados meses do ano de previsão fazendo com que o modelo de regressão cometa maior erro.

Considerando que houve uma mudança na variabilidade 'natural' desta série de dados (aumento das medidas nos anos iniciais, devido ao enchimento do reservatório e posterior estabilização), as principais contribuições do trabalho são a determinação do período de maior acurácia da previsão e de um modelo com boa precisão, em média, 6%.

Agradecimentos

Ao CEASB -Centro de Estudos Avançados em Segurança de Barragem pela cessão dos dados e as bolsas concedidas.

Referências

- [1] Box, G. E. P., Jenkins, G. M. e Reinsel, G. C. *Time Series Analysis Forecasting and Control - Fourth Edition*. Hoboken-NJ: John Wiley and Sons, Inc., 2008.
- [2] Fusaro, T. C. *Estabelecimento estatístico de valores de controle para instrumentação de barragens de terra - Estudo de caso das barragens de Emborcação e Piau*. Ouro Preto: Dissertação de Mestrado Profissional em Engenharia Geotécnicas da UFPOF, 2007.

- [3] ICOLD - CIGB. *Dams & The World's Water - An educational book that explains how dams help to manage the world's water*. Paris: International Commission on Large Dams/Comission Internationale des Grand Barrages, 2008.
- [4] Itaipu Binacional. *Síntese das Principais Informações de Projeto para o Controle da Instrumentação da Barragem Principal - Trechos E, F e I* (Relatório 4106.50.06001). Foz do Iguaçu, 1984.
- [5] Montgomery, D. C. *Introdução ao Controle Estatístico de Qualidade, 7ª edição*. Rio de Janeiro: LTC, 2019.
- [6] Morettin, P. A. *Análise de séries temporais, 2ª edição rev. e ampl.*. São Paulo: Edgard Blucher, 2006.
- [7] Salazar, F. *et al.* Data-Based Models for the Prediction of Dam Behaviour: A Review and Some Methodological Considerations. *Archives of Computational Methods in Engineering*, p. 1134-1160. DOI 10.1007/s11831-015-9157-9.
- [8] Silveira, J. F. A. *Instrumentação e comportamento de fundações de barragens de concreto*. São Paulo: Oficina de Textos, 2003.
- [9] USACE. *Enginnering and Design - Instrumentation of Embankment Dams and Levees*. U.S.Army Corps of Engineers. Washington, DC. 1995.