

Impacto dos *outliers* na modelagem de um índice estatístico

Impact of outliers on modeling a statistical index

RESUMO

Leonardo Oro Lucini
leonardooroutf@gmail.com
Universidade Tecnológica Federal do Paraná, Francisco Beltrão, Paraná, Brasil

Sheila Regina Oro
sheilaro@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Paraná, Brasil

Luiz Fernando Zubeldia Dal Posso
luiz.zubeldia@pti.org.br
Parque Tecnológico Itaipu, Foz do Iguaçu, Paraná, Brasil

Claudio Issamy Osako
osako@itaipu.gov.br
Itaipu Binacional, Foz do Iguaçu, Paraná, Brasil

O tema da segurança de barragens tem se mostrado muito importante atualmente. Diversas pesquisas têm utilizado métodos estatísticos que, aplicados junto com as técnicas de monitoramento, visam avaliar a confiabilidade da estrutura. O Índice de Monitoramento Conjunto das Respostas dos Blocos da Barragem foi obtido neste estudo, com base nas informações provenientes das leituras dos instrumentos de segurança contidos no bloco chave F35/36 da barragem principal de Itaipu. O objetivo principal foi avaliar o impacto de uma das etapas da modelagem deste índice, a qual se refere pela adequação de valores atípicos (*outliers*). Estes valores que se não tratados impactam negativamente a capacidade de representação do modelo, fato que se comprovou ao final do estudo.

PALAVRAS-CHAVE: Monitoramento de barragens. Análise Fatorial. Bloco Chave.

ABSTRACT

The topic of dam safety has proved to be very important today. Several researches have used statistical methods that, applied together with the monitoring techniques, aim to assess the reliability of the structure. The Joint Monitoring Index of the Dam Blocks was formulated in this study, based on information from security instruments readings contained in key block F35 / 36 of the main Itaipu dam. The main objective was evaluate the impact of one of the stages while modeling this index, which is referred to by adequacy of outliers. These values that, if not treated, influence negatively the ability to represent the model, a fact that was proven at the end of the study.

KEYWORDS: Monitoring of dams. Factor Analysis. Key Block.

Recebido: 19 ago. 2020.

Aprovado: 01 out. 2020.

Direito autorial: Este trabalho está licenciado sob os termos da Licença Creative Commons-Atribuição 4.0 Internacional.



INTRODUÇÃO

Toda estrutura construída transversalmente a um rio ou talvegue que tem como objetivo elevar o seu nível d' água e/ou de criar um reservatório de acumulação de água seja de regulação das vazões do rio, seja de outro fluido recebe o termo de Barragem (MSIB, 2002, p.15). Por se tratar de uma obra de grande porte, essa requer extremos cuidados em relação a sua segurança. Segundo o Manual de Segurança e Inspeção de Barragens do Ministério da Integração Nacional da Agência Nacional de Águas (ANA, 2002, p.18), “uma barragem segura é aquela cujo desempenho satisfaça as exigências de comportamento necessárias para evitar incidentes e acidentes que se referem a aspectos estruturais, econômicos, ambientais e sociais”.

A instrumentação de barragens aliada às técnicas estatísticas, nos permitem compreender o comportamento da barragem ao longo do tempo. A Análise Fatorial, que é fruto da Análise Multivariada, é uma dessas técnicas que podem ser aplicadas (HAIR et al, 2009). Além disso, é com base nessa técnica em que foi desenvolvido o Índice de Monitoramento Conjunto das Respostas dos Blocos da Barragem (IMCRB), que foi criado para representar o conjunto de observações dos sensores utilizados em um mesmo estudo, reduzindo ao máximo a perda de informação (ORO, 2016).

Na análise fatorial, explica-se todas as covariâncias, utilizando-se poucas variáveis não observáveis, que são denominados fatores comuns, desde que seja possível modelar. Segundo Ferreira (2018, p. 460), “os fatores comuns são responsáveis pela explicação tanto das variâncias quanto das covariâncias, enquanto as variáveis específicas contribuem somente para as variâncias das variáveis originais consideradas no modelo”. De acordo Silva, Marques e Nieradka (2015), “a análise fatorial possibilita o tratamento das leituras dos instrumentos como variáveis aleatórias, permitindo gerar fatores altamente correlacionados e encontrar relações entre eles que não são facilmente observáveis”.

Segundo Martens (2014), “um problema recorrente em dados de séries temporais é a presença de erros grosseiros e influências externas, como por exemplo, enchentes, greves, e mudanças políticas e econômicas”. Quando se trata de barragens, surgem novos exemplos, entre eles estão variação de temperatura, processos de drenagem, entre outros. O autor também afirma que “estes fatos, comumente chamados de outliers, representam uma preocupação quanto a eficiência das estimativas dos parâmetros dos modelos e, conseqüentemente, a previsão de observações futuras”.

Portanto, com base nas observações realizadas pelos instrumentos do bloco chave F35/36 presente na Barragem Principal de Itaipu, o objetivo principal deste estudo foi realizar um comparativo entre duas modelagens do IMCRB. A primeira foi realizada com os dados brutos fornecidos, enquanto que a segunda se optou por adequar os *outliers* encontrados, etapa que se provou essencial no fim deste estudo.

MATERIAL E MÉTODOS

A Usina Hidrelétrica de Itaipu está localizada no rio Paraná, na divisa entre o Brasil e o Paraguai, 14 km a montante da Ponte Internacional da Amizade. Os dados

utilizados neste estudo foram fornecidos pelo Centro de Estudos Avançados em Segurança de Barragens (Ceasb), e referem-se às medições manuais mensais, realizadas entre janeiro de 2001 e dezembro de 2019, de 39 instrumentos presentes no bloco chave F35/36, localizado na barragem principal de Itaipu. Os instrumentos considerados foram Pêndulos Direto e Invertido, Piezômetros, Extensômetros Múltiplos de Hastes, Bases de Alongômetro e Medidores Triortogonais.

A primeira etapa, de organização e tabulação dos dados, foi realizada com o apoio de planilhas eletrônicas. Nela, notou-se que em alguns meses não haviam leituras, portanto seria necessário realizar o preenchimento destas lacunas. Para isso, contou-se com o auxílio do *software Matlab* (2020), onde os dados foram processados até obter uma matriz de médias mensais completa. Na sequência, foi feita a representação dos instrumentos por um número menor de variáveis por meio da aplicação da análise fatorial aos dados padronizados, implementando-se no *software R* (2020) a sintaxe apresentada na Figura 1.

Figura 1- Sintaxe da Análise Fatorial no *software R*

```
library(psych)
af<-fa(dados, nfactors=k, fm = "ml", rotate = "varimax")
summary(af)
```

Fonte: Autoria própria (2020).

Em seguida, foi investigado no conjunto de leituras dos instrumentos, a presença de valores atípicos quando comparados aos de sua vizinhança (*outliers*). Esta ação foi realizada também no R, com a sintaxe apresentada na Figura 2, a qual também permitiu realizar uma previsão de qual seria um valor aproximado para aquela leitura. A matriz de dados foi atualizada com os *outliers* sendo substituídos pelos valores previstos e novamente realizou-se a análise fatorial.

Figura 2- Sintaxe *Outliers* no *software R*

```
library(forecast)
dados <- ts(pdx[,1:30], start=c(2001, 1), frequency=12)
plot(dados[,20])
tsoutliers(dados[,2], iterate = 0, lambda=NULL)
dados[101,44]
```

Fonte: Autoria própria (2020).

Com os resultados de cada análise fatorial, foi modelado o Índice de Monitoramento Conjunto das Respostas dos Blocos da Barragem (IMCRB), como uma combinação linear dos escores fatoriais, com os pesos obtidos com base nos autovalores da matriz de correlação dos dados, conforme a Eq. (1). Utilizou-se este método para ambas as modelagens realizadas, com exceção da etapa de adequação dos *outliers* para a primeira, a qual trabalhou-se com os dados brutos fornecidos.

$$\text{IMCRB} = C1 * \text{fatores} [, 1] + C2 * \text{fatores} [, 2] + C3 * \text{fatores} [, 3] \dots + Cn * \text{fatores} [, n] \quad (1)$$

RESULTADOS E DISCUSSÃO

Foram detectados 133 *outliers* distribuídos em 29 instrumentos. Dentre eles, o instrumento que mais apresentou esses valores foi a Base de Alongâmetro x28, apresentando 18 valores atípicos.

A determinação de fatores resultou-se por meio da Análise Paralela de Horn e do Critério de Kaiser, nos quais, em um primeiro momento, foram obtidos 5 e 6 fatores respectivamente, e após a substituição dos valores atípicos este número se manteve em 5 para ambos os casos. Essa melhora é de grande importância, uma vez que buscamos representar o conjunto de dados com o menor número de fatores possíveis, a fim de evitar o acúmulo de erros durante todo o processo de modelagem. É importante ressaltar que as análises feitas a seguir, levaram em conta a comparação das análises fatoriais com 5 fatores, uma vez que não havia grande melhora entre 5 e 6 fatores citados anteriormente.

Com a adequação dos *outliers*, pode-se observar um impacto positivo no grau de explicação do modelo fatorial. Após esta ação, foi obtido um resultado em torno de 79,1%, enquanto que antes a mesma resultava em 75,15%. Com isso, observa-se um acréscimo de aproximadamente 4% na capacidade de representação das informações dos instrumentos por grupo de fatores.

Outra característica impactada pela adequação dos valores atípicos foram as comunalidades dos instrumentos, que representam o quanto os instrumentos estão sendo bem explicados pelo modelo, e, portanto, é preferível que resultem maior que 0,5. Dentre os 39 instrumentos estudados, 7 apresentaram comunalidades baixas na análise em que os valores atípicos não foram ajustados. Em contrapartida, quando houve a adequação desses valores, apenas 5 instrumentos apresentaram este problema. Do total de instrumentos presentes no estudo, 21 obtiveram melhoras, número que representa 53,46% do conjunto. A Tabela 1 apresenta os instrumentos que sofreram um maior impacto, seja positivo ou negativo. É importante ressaltar o impacto positivo nos Extensômetros x15 e x16, estes que praticamente tiveram a sua representatividade dobrada.

Tabela 1 - Instrumentos x Comunalidades

Instrumentos	Comunalidade (Com Outliers)	Comunalidade (Sem Outliers)
x1	0,12	0,30
x3	0,84	0,91
x4	0,76	0,85
x5	0,71	0,80
x6	0,40	0,46
x7	0,83	0,76
x9	0,99	0,87
x10	0,89	0,86
x11	0,99	0,88
x15	0,50	0,97
x16	0,44	0,91
x21	0,66	0,69

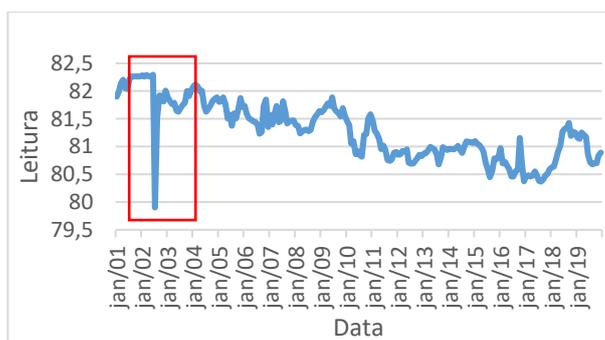
Instrumentos	Comunalidade (Com Outliers)	Comunalidade (Sem Outliers)
x31	0,90	0,95
x38	0,92	0,96

Fonte: Autorial Própria (2020)

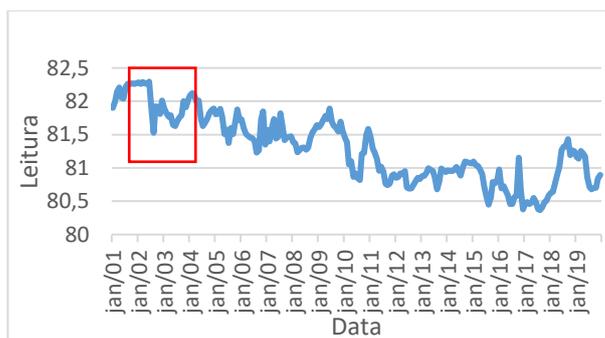
Em relação às cargas fatoriais, característica esta que se refere a representatividade do instrumento em relação aos fatores, não houve impacto significativo da etapa de substituição dos *outliers*. Os instrumentos que antes melhor representavam os fatores, permanecem representando os mesmos após esta etapa.

Visualmente, esta melhoria pode ser observada ao compararmos os gráficos dos instrumentos após a etapa de substituição dos valores atípicos. Como por exemplo no caso do piezômetro x4, em que a Figura 3 (a) refere-se à série temporal original, enquanto que a Figura 3 (b) refere-se àquela após a substituição dos outliers. Essa melhora que pode ser expressada em aproximadamente 9% de representatividade em relação ao modelo.

Figura 3 – Série temporal do piezômetro x4



(a) Dados originais

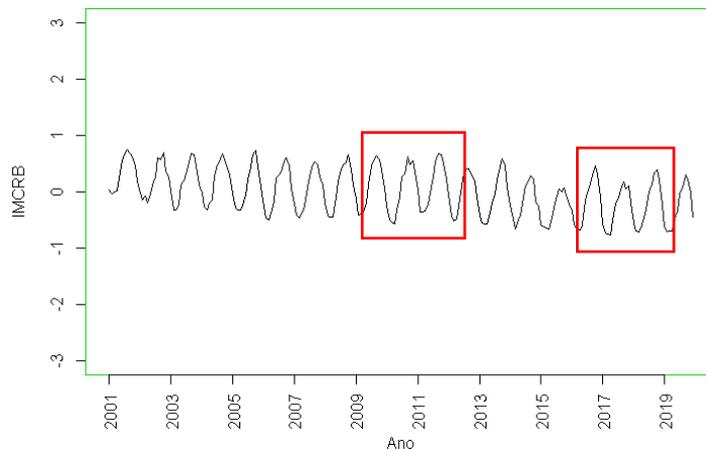


(b) Após substituição do outlier

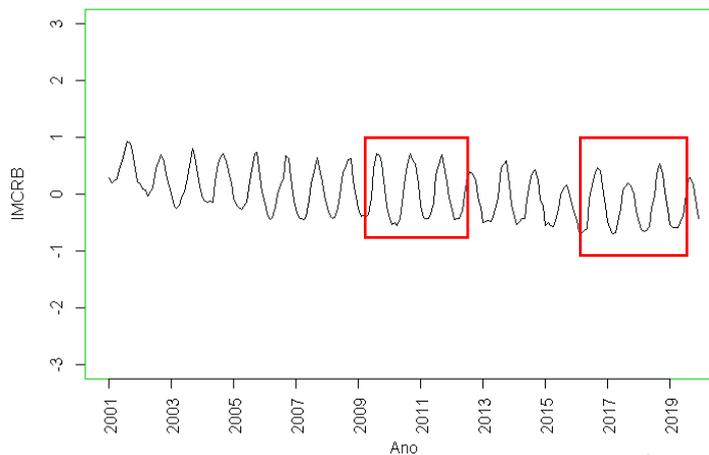
Fonte: Autorial própria (2020).

Por fim, é possível observar esses impactos no gráfico do IMCRB modelado. É perceptível que o comportamento dos gráficos, ilustrados na Figura 4 (a) e 4 (b) se manteve o mesmo. Porém, houve uma suavização dos picos de máximo e mínimo referentes a sazonalidade presente no gráfico após a substituição dos *outliers*.

Figura 4 - IMCRB Bloco F35/36



(a) Dados originais



(b) Após substituição dos outliers

Fonte: Autoria própria (2020).

CONCLUSÃO

O método aplicado neste estudo é um complemento para o monitoramento de risco em um bloco de concreto, composto por um grande número de instrumentos.

A adequação dos *outliers*, permitiu aumentar a correlação dos instrumentos e também a capacidade de representação destes por meio dos fatores. Uma vez que esses valores foram ajustados, os resultados foram observados imediatamente, tanto visualmente a partir dos gráficos dos instrumentos, como também numericamente nas informações provenientes da análise fatorial. Portanto, é possível afirmar que a substituição dos *outliers* é de grande importância para a modelagem do IMCRB.

AGRADECIMENTOS

O presente trabalho foi realizado com o apoio do Parque Tecnológico Itaipu (PTI), do Centro de Estudos Avançados em Segurança de Barragens (CEASB) e da Universidade Tecnológica Federal do Paraná/Brasil.

REFERÊNCIAS

FERREIRA, D. F. **Estatística multivariada**. 3. ed. Lavras: Ed. UFLA, 2018. 624p.

HAIR, J. F. et al. **Análise multivariada de dados**. 6. ed. Porto Alegre, RS: Bookman, 2009. 688 p.

MARTENS, M. **Identificação e avaliação do impacto de outliers em processos ARFIMA (p, d, q)**. 2014. 79f. Trabalho de Conclusão de Curso (Bacharelado em Estatística).

MSIB. **Manual de Segurança e Inspeção de Barragens**. Brasília: Ministério da Integração Nacional, 2002. 148p.

ORO, S. R. **Índice de monitoramento do comportamento estrutural dos blocos de concreto de barragens: uma abordagem multivariada**. 2016. 139f. Tese (Doutorado em Métodos Numéricos em Engenharia) – Universidade Federal do Paraná, Curitiba, 2016.

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

SILVA, Orlando Catarino; MARQUES, Jair Mendes; NIERADKA, Itamar Pena. **Testes Estatísticos e Análise Multivariada na Determinação dos Modos de Falha Monitorados Pelos Instrumentos em Blocos de Concreto da Barragem de Itaipu**. HOLOS, v. 5, p. 3-21, 2015.

The Mathworks, Inc., Natick, Massachusetts, 9.8.0.1359463 (R2020a) Update 1 - License Number: STUDENT, 2020.