Revista Contexto GEOGRÁFICO ISSN 2595-7236
Universidade Federal de Alagoas - Instituto de Geografia, Desenvolvimento e Meio Ambiente
Programa de Pós-Graduação em Geografia
http://www.seer.ufal.br/index.php/contextogeografico
Recebido em: 15//08/2023
Aceito para publicação em: 10/09/2023

DESEMPENHO DO MÉTODO ESTATÍSTICO DE REGRESSÃO LINEAR MÚLTIPLA NO PREENCHIMENTO DE FALHAS EM DADOS PLUVIOMÉTRICOS

Luis Felipe Santos Moura
Universidade Federal do Ceará – UFC, Brasil
felipesantos010186@gmail.com

Carlos Henrique Sopchaki
Universidade Federal do Ceará – UFC, Brasil
carlos.geografia@ufc.br

RESUMO – Esta pesquisa tem como objetivo aplicar o método estatístico de regressão linear múltipla no preenchimento de falhas para uma região do semiárido e com isso observar o seu desempenho conforme cinco intervalos de faixas de falhas. Para tanto, o trabalho utilizou-se de uma série histórica de 59 anos disponibilizada pela Agência Nacional de Águas, onde, através do processo de regionalização escolheu-se outras três estações próximas para estimar os dados da estação com falha através de modelo de regressão linear múltipla. Após tratamento e interpretação dos resultados, foi observado que a faixa de falha na ordem de 20% apresentou o maior coeficiente de variação de erros relativos, enquanto que na faixa de 30% de falhas houve menor índice de erros relativos. Porém analisando a qualidade (avaliação de desempenho) dos dados, observou-se que os mesmos demonstraram que a estimação apresenta uma baixa confiabilidade, apesar de sua vantagem no baixo desvio entre os dados reais e estimados.

Palavras-chave: Dados de precipitação; Modelagem estatística; Microsoft Excel.

PERFORMANCE OF THE STATISTICAL METHOD OF MULTIPLE LINEAR REGRESSION IN FILLINGS GAPS IN RAINFALL DATA

ABSTRACT – This research aims to apply the statistical method of multiple linear regression to fill faults in a semi-arid region and thus observe its performance according to 5 intervals of fault bands. Therefore, the work used a 59-year historical series available by "Agência Nacional de Águas", which, through the regionalization process, three other nearby stations were chosen to estimate the data from the failed station using a multiple linear regression model. After processing and interpreting the results, it was observed that the fault band in the order of 20% presented the highest coefficient of variation of relative errors, while in the failure range of 30% there was a lower rate of relative errors. However, analyzing the quality (performance evaluation) of the data, it was observed that they demonstrated that the estimation has low reliability, despite its advantage in the low deviation between real and estimated data.

Keywords: Precipitation data; Statistical modeling; Microsoft Excel.

INTRODUÇÃO

Dados pluviométricos são extremamente importantes nos mais diversos aspectos das ciências ambientais. A análise de dados pluviométricos é importante para tratativas relativas a, por exemplo, riscos a eventos hidrológicos, mudanças climáticas, gestão de recursos hídricos e planejamento estratégico, fornecendo primordialmente, dados primários para os interessados.

Zarekarizi et al. (2016) complementam citando que dados de séries históricas ajudam a compreender a complexidade dos fenômenos relativos à precipitação, tanto para verificação de médias históricas quanto para espacialização e modelagem de eventos futuros. Estudos com estes vieses ambientais, que demandem de simulação e de variabilidade climática, principalmente à longo prazo, como por exemplo dimensionamento de barragens, mudanças climáticas ou clima urbano, necessitam de séries pluvio e fluviométricas temporais confiáveis (LI et al., 2010; PAZ, 2010).

Entretanto, muitos destes dados estão sujeitos a falhas pois pode ocorrer, por exemplo, omissão da entrega dos dados por parte do catalogador, problemas de acesso à estação meteorológica ou danificação do equipamento, que podem levar a observações erradas, incompletas ou até mesmo a completa ausência de dados (MWALE, ADELOYE e ROSTUM, 2012; BIER e FERRAZ, 2017).

A utilização de séries históricas pluviométricas com erros ou com falhas pode levar a análises equivocadas por parte dos pesquisadores. Em simulação de cenários hidrológicos pode levar a dificuldades nos ajustes de distribuições estatísticas dos dados, reduzir o desempenho ou inviabilizar a aplicabilidade de modelagem em hidrologia e acarretar na publicação de informações errôneas à sociedade. Dessa forma, a observação de falhas nos dados deve ser obrigatória antes de trabalhar em dados de precipitação e verificar a sua consistência (DEPINÉ et al., 2014).

No estado do Ceará, o monitoramento dos dados pluviométricos é feito pela Fundação Cearense de Meteorologia e Recursos Hídricos (FUNCEME) e pela Agência Nacional de Águas (ANA). Esta última, além de realizar o processamento dos dados pluviométricos e fluviométricos, também desenvolve alertas contra cheias em áreas críticas. Assim, tornando relevante a presença e acurácia em seus dados meteorológicos. Na impossibilidade do catálogo de dados pluviométricos completo, existem diversos métodos que realizam o preenchimento de falhas em séries históricas.

Quando não é possível a coleta pontual de um dado pluviométrico, uma das possibilidades de simular seu valor próximo é valer-se de métodos de estimação estatística para espacializar de forma aproximada as chuvas que ocorreram (CECÍLIO e PRUSKI, 2003). Dentre os métodos existentes pode-se citar: razão normal, ponderação pelo inverso da distância, regressão linear simples (RLS) ou múltipla (RLM), redes neurais artificiais, vetor regional, regressão linear com peso R ou ponderação regional.

Em relação ao método de regressão, a múltipla diferencia-se da simples no fato de que enquanto a simples avalia a relação entre duas variáveis, a múltipla considera uma relação de mais de duas variáveis. A RLM é bastante difundida pela literatura acadêmica pois é um método considerado fácil e simples de predizer os dados, tanto por conta de sua versatilidade de modelagem pelo geoprocessamento ou em softwares estatísticos, quanto no seu cálculo manual (KHOSRAVI et al., 2015; BIER e FERRAZ, 2017; MILOVANOVIĆ et al., 2017; MELLO, KOHLS e OLIVEIRA, 2017). Na hidrologia estes métodos também podem ser importantes aliados no processo de estimação através de regionalização de dados hidrológicos.

Apesar de compreender a sua importância, a RLM, assim como qualquer outro método estimativo também pode ter suas desvantagens e limitações para operação. Considerando isso, também é importante que se estude os seus padrões, observando em que momento as quantidades de falhas existentes em uma série melhor ou pior estimam em relação aos dados reais. Assim como observar se o próprio método possui um desempenho estatístico aceitável para preencher quaisquer dados.

Diante disso, o objetivo deste trabalho é realizar o preenchimento de falhas em dados pluviométricos, utilizando-se do método de regressão linear múltipla (RLM). Para tanto, serão criadas falhas fictícias aleatórias para a área de ensaio do posto pluviométrico de Piquet

Carneiro (CE). O método utilizado tem como base a estimação por regionalização hidrológica, utilizando-se de dados de três postos pluviométricos vizinhos com dados completos, sendo eles Mombaça, Milhã e Senador Pompeu.

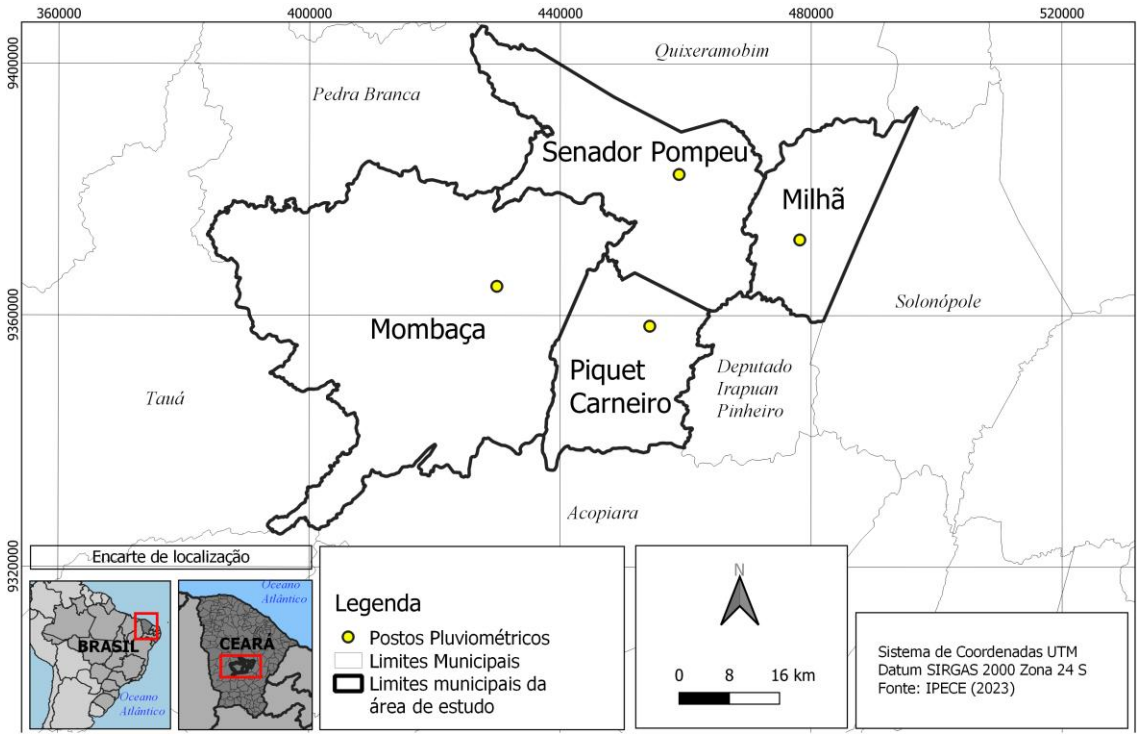
MATERIAIS E MÉTODOS

Caracterização da área de estudo

A área de estudo (figura 1) compreende quatro municípios que se localizam na região do Sertão Central do estado do Ceará. O regime climático preponderante é semiárido com precipitação média anual entre 720 e 780 milímetros (mm) anuais. Assim como todo semiárido brasileiro, existe uma grande variabilidade climática condicionada pela atuação de vários sistemas atmosféricos provocadores de chuva concentrados no primeiro semestre do ano (MARENGO *et al*., 2011).

Analisando dados hipsométricos e de declividade (VALERIANO e ROSSETI, 2011), o relevo na região de pesquisa é preponderantemente plano, sob uma cota média de 200 metros de altitude, assim, observa-se que a região não sofre influência de aspectos orográficos na pluviometria. Estas observações são bastante importantes para a regionalização hidrológica ao contribuírem na identificação de zonas pluviometricamente homogêneas.

Figura 1. Localização da área de pesquisa.



Org. Os autores (2023).

Dados pluviométricos e regionalização hidrológica

Os dados pluviométricos foram adquiridos através da plataforma *Hidroweb* da Agência Nacional de Águas (ANA). Em seguida, os dados foram tratados através do *Software* Microsoft Excel escolhendo-se a escala temporal anual. A escolha da escala se deveu ao fato da alta variabilidade pluviométrica e disponibilidade de dados, pois caso fosse escolhido média mensal

seria necessário considerar meses que possuem dados próximos de 0 mm, que ocorrem na estação seca. Para dados diários, Fill (1987) compara várias metodologias de preenchimento de falhas e comenta que nenhuma é adequada para o preenchimento de falhas diárias, sendo mais recomendadas no preenchimento de falhas mensais ou anuais.

Para corrigir as falhas, tendo como área de ensaio o posto pluviométrico de Piquet Carneiro, utilizou-se de outros três postos vizinhos da área de pesquisa. No processo conhecido como regionalização hidrológica, o único requisito é de que haja correlação no comportamento das chuvas nos postos com dados normais (TUCCI, 2002; BERTONI e TUCCI, 2007), ou seja, deve ser observado se as mesmas se localizam em zonas pluviometricamente homogêneas podendo ser analisado de forma simples, através do paralelismo entre os dados de chuva média mensal, declividade e altimetria, dessa forma, há uma maior confiabilidade na correlação da estimativa sem interferência de fatores ambientais.

Rodrigues (2002) define a regionalização hidrológica como um conjunto de procedimentos e métodos estatísticos que visam explorar ao máximo os dados existentes numa região hidrológica, buscando-se permitir a estimativa da vazão ou pluviometria desejada num local com ausência de dados ou com dados muito escassos.

Modelagem estatística e tratamento de falhas nos dados pluviométricos

Para que se possa aplicar um método de preenchimento de falhas em dados pluviométricos recomenda-se primordialmente que se faça um teste através do diagrama de *Gantt* (SAF, 2010). O intuito deste diagrama é demonstrar a disponibilidade temporal dos dados, assim, por exemplo, estações com séries de mais de N dias ou N% de falhas nos dados pluviométricos no ano são consideradas inaptas para serem utilizados e são consideradas como falhas.

Porém, para que se possa observar o desempenho do método, decidiu-se simular anos com falhas ao inserir lacunas propositais em anos da série histórica de Piquet Carneiro. Deste modo, ocultou-se o registro normal da série, deixando os anos com falhas, ou seja, com as células vazias. Porém, o registro pluviométrico normal também é utilizado para fazer posteriores comparações entre os dados reais e os estimados.

A quantidade de falhas fictícias alocadas varia nas faixas de 10%, 20%, 30%, 40% e 50% dos dados pluviométricos da série histórica de 59 anos (1949-2018).

A escolha dos limiares das faixas de falhas tem o intuito de observar o desempenho do método de estimação ao longo das faixas de falhas que será observada através da análise de erros relativos (desvio) e coeficiente de variação.

A alocação de anos com falhas nos dados foi indicada de maneira aleatória através de programação Python utilizando um script com a função *Numpy Array*, que gera um número e posição aleatória de dados que devem ser ocultados conforme o percentual solicitado da faixa falhas escolhido. O intuito deste processo é evitar o enviesamento das falhas.

Assim, se há 59 anos de série histórica, cerca de 5,9 anos ≈ seis anos estarão com falhas para o intervalo de 10%, por exemplo. Utilizando deste *script*, as falhas aleatórias foram aplicadas para cada faixa de falha que está exemplificada de forma parcial (1949-1980) no quadro 1, onde as células que estão destacadas em bege são os anos escolhidos para a inserção de lacunas na série histórica através do software Excel.

Quadro 1. Alocação de falhas aleatórias para uma parte da série histórica.

FAIXAS	10%			20%			30%			40%			50%		
Ano	SF *	SO **	RLM ***	SF	SO	RLM	SF	SO	RLM	SF	SO	RLM	SF	SO	RLM
1949	495,8	495,8		495,8	495,8		495,8	495,8		495,8	495,8		X	495,8	482,9
1950	565,3	565,3		565,3	565,3		565,3	565,3		565,3	565,3		X	565,3	572,1
1952	521,9	521,9		521,9	521,9		521,9	521,9			521,9	558,0	X	521,9	513,1
1953	388,2	388,2		388,2	388,2		388,2	388,2		388,2	388,2		388,2	388,2	
1954	389,4	389,4		389,4	389,4		X	389,4	378,0	389,4	389,4		389,4	389,4	
1955	323,9	323,9		323,9	323,9		323,9	323,9		323,9	323,9		X	323,9	343,6
1962	640,0	640,0		X	640,0	486	640,0	640,0		X	640,0	652,2	640,0	640,0	
1963	784,3	784,3		784,3	784,3		X	784,3	776,6	X	784,3	799,8	X	784,3	770,0
1964	925,9	925,9		925,9	925,9		925,9	925,9		X	925,9	1011,0	X	925,9	983,6
1965	625,4	625,4		625,4	625,4		625,4	625,4		X	625,4	666,8	625,4	625,4	
1966	511,3	511,3		511,3	511,3		511,3	511,3		511,3	511,3		511,3	511,3	
1967	925,6	925,6		X	925,6	1645,5	925,6	925,6		X	925,6	928,0	X	925,6	908,0
1968	X	598,9	715,6	598,9	598,9		598,9	598,9		598,9	598,9		X	598,9	715,6
1969	747,5	747,5		747,5	747,5		747,5	747,5		747,5	747,5		747,5	747,5	
1970	333,4	333,4		333,4	333,4		X	333,4	215,9	333,4	333,4		X	333,4	336,1
1971	557,1	557,1		557,1	557,1		557,1	557,1		X	557,1	574,5	557,1	557,1	
1972	252,8	252,8		252,8	252,8		X	252,8	259,0	252,8	252,8		252,8	252,8	
1974	1203,9	1203,9		1203,9	1203,9		X	1203,9	1229,6	X	1203,9	1297,1	X	1203,9	1223,0
1975	728,1	728,1		728,1	728,1		728,1	728,1		X	728,1	743,0	728,1	728,1	
1976	532,1	532,1		532,1	532,1		532,1	532,1		532,1	532,1		532,1	532,1	
1977	640,9	640,9		640,9	640,9		640,9	640,9		X	640,9	623,7	640,9	640,9	
1978	362,5	362,5		362,5	362,5		362,5	362,5		X	362,5	478,6	X	362,5	458,4
1980	563,5	563,5		X	563,5	537,5	563,5	563,5		563,5	563,5		X	563,5	554,2

Org. Os autores (2023). Notas: * Série Histórica com falhas; ** Série histórica original; *** Dado estimado por RLM.

Com as falhas alocadas, o passo posterior é ajustar os dados para estimá-los por meio do modelo linear utilizando-se da RLM (equação 1). A aplicação do método é possível dentro do *software* Microsoft Excel no qual, após selecionar as colunas de variável dependente (Y) e variáveis independentes (X), o programa fornece os dados de estatísticas de regressão e análise de variância (ANOVA).

$$Y_i = \alpha + \beta_1 X_{1_i} + \beta_2 X_{2_i} + \ldots + \beta_k X_{k_i} + e_i \quad (1)$$

Onde: α é o valor esperado de Y quando todos as variáveis independentes forem nulas; β_1 é a variação esperada em Y dado um incremento unitário em X1 (primeira variável), mantendo-se constantes todas as demais variáveis independentes; β_k é a variação esperada em Y dado um incremento unitário em Xk (k variáveis), mantendo-se constantes todas as demais variáveis independentes; e_i é o erro não explicado pelo modelo.

A ferramenta de análise de dados por meio da RLM através do Microsoft Excel fornece bases que podem ser usados para desenvolver essa equação e também outros, como:

r-múltiplo – refere-se à correlação entre os dados de chuva na estação com problema e aquele que foi estimado;

r-quadrado – também chamado de coeficiente de determinação, indica o quanto de percentual da chuva que acontece na região pode ser informada pela regressão;

erro padrão – indica que a regressão erra em média X mm de chuva;

coeficientes – É o valor da variável independente caso todas as outras variáveis sejam hipoteticamente zeradas. Existe um valor para cada variável (postos com dados) e também para a interseção (regionalização entre os três postos para a estimação);

valor-p – derivado da estatística do teste de hipóteses, indica a probabilidade de se obter este resultado ao acaso, quanto menor, mais significativo. Comumente resultados abaixo de 5% ou 0,05 indicam significância ideal;

resíduos – Indica a diferença entre o dado estimado e o registrado. Quanto mais distante da reta de regressão, maior a diferença, ou seja, quanto maior ou menor os resíduos, mais díspares são os resultados da estimação.

Para que se possa preencher as falhas utilizando estes dados faz-se necessário desenvolver um *script* dentro da célula com falha. O *script* está elencado na equação 2 para o exemplo de falhas aplicadas em uma célula com falha de 10% da série histórica em duas pastas de trabalho diferentes, uma com falha nos dados da série e outra com as análises das estatísticas de regressão.

='Analise_10%'!I19'10%'!B14+'Analise_10%'!I20*'10%'!C14+'Analise_10%'!I21*'10%'!D14 +'Analise_10%'!I18* (2)

A equação 2 pode ser descrita facilmente quando comparada com a equação de regressão (equação 1) ou descrevendo cada parte, a dizer:

'Analise_10%'!I19 [...] 'Analise_10%'!I20 [...] 'Analise_10%'!I21: referem-se aos dados dos coeficientes obtidos através dos dados da análise de regressão para os postos com dados completos usados na regionalização (Mombaça, Senador Pompeu e Milhã);

'10%'!B14 [...] '10%'!C14 [...] '10%'!D14: referem-se aos dados de chuva registrada na pasta que contém os dados de série histórica para os postos com dados completos;

'Analise_10%'!I18: refere-se ao dado de coeficiente de interseção da pasta que contém os dados da análise de regressão.

Após aplicar a função na célula vazia (com falha), foi gerado um dado pluviométrico estimado, tal qual como o registrado na coluna "RLM" do quadro 1. A função foi aplicada às células

vazias e em cada pasta de trabalho com intervalos os intervalos de falhas de 10%, 20%, 30%, 40% e 50%.

Com os dados pluviométricos estimados foi possível elaborar gráficos e *boxplots* comparativos entre os dois dados de forma simplificada, como também análises mais detalhadas como erros relativos, menor erro, maior erro, média de erro, desvio padrão dos erros e coeficiente de variação do erro, possibilitando verificar os desvios entre os dados estimados e os dados reais.

Avaliação de análise de desempenho das estimativas

Para analisar se o ajuste dos dados gerados das séries originais está dentro de um valor estatisticamente confiável em relação aos dados ajustados pelo modelo de RLM, decidiu-se calcular critérios de avaliação de desempenho ou avaliação da qualidade da relação linear entre os dados.

Conforme Anthes *et al*. (2006), a análise de desempenho das estimativas tem o intuito de dimensionar os erros nas simulações numéricas calculadas e a possível indicação das suas fontes de erros, aplicando-se uma série de índices e escores estatísticos, os quais são usados como ferramentas de avaliação da acurácia dos experimentos numéricos.

Para tanto utilizou-se quatro critérios de avaliação de desempenho dos dados: Erro Médio Absoluto (Mean Absolute Error - MAE), conforme a equação 3; Raiz Quadrada do Erro Médio (Root Mean Square Error - RMSE), conforme a equação 4; Índice de Destreza de Scores (Skill Score - SS), conforme a equação 5; Coeficiente de Correlação de Pearson (Pearson's Correlation Coefficient - PCC), conforme a equação 6.

$$MAE = \frac{\sum_{i=1}^{n} |S_{oi} - S_{ei}|}{n} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (S_{oi} - S_{ei})^2}{n}} \quad (4)$$

$$SS = 1 - \frac{S_{oi}}{S_{ei}} \quad (5)$$

$$PCC = \frac{\sum_{i=1}^{n} (S_{oi} - \overline{S_o})(S_{ei} - \overline{S_e})}{\sqrt{\sum_{i=1}^{n} (S_{oi} - \bar{S})^2 \sum_{i=1}^{n} (S_{ei} - \bar{S})^2}} \quad (6)$$

Onde: S_{oi} representa o valor observado de precipitação no ano i; S_{ei} representa o valor estimado pela modelagem de RLM no ano i; n representa o número de dados utilizados no processo; e $\overline{S_o}$ e $\overline{S_e}$ são respectivamente a média de valores observados de precipitação e a média de valores estimados de precipitação.

RESULTADOS E DISCUSSÃO

Análise de desempenho do método de estimação

Observando os dados de distribuição acumulada da figura 2 que analisa a amplitude geral dos dados, observa-se que o modelo de RLM teve um ajuste aos dados com pouco desvio entre os dados normais e aqueles estimados em sua integralidade, demonstrando que no geral há uma boa homogeneidade dos dados, com poucos *outliers*.

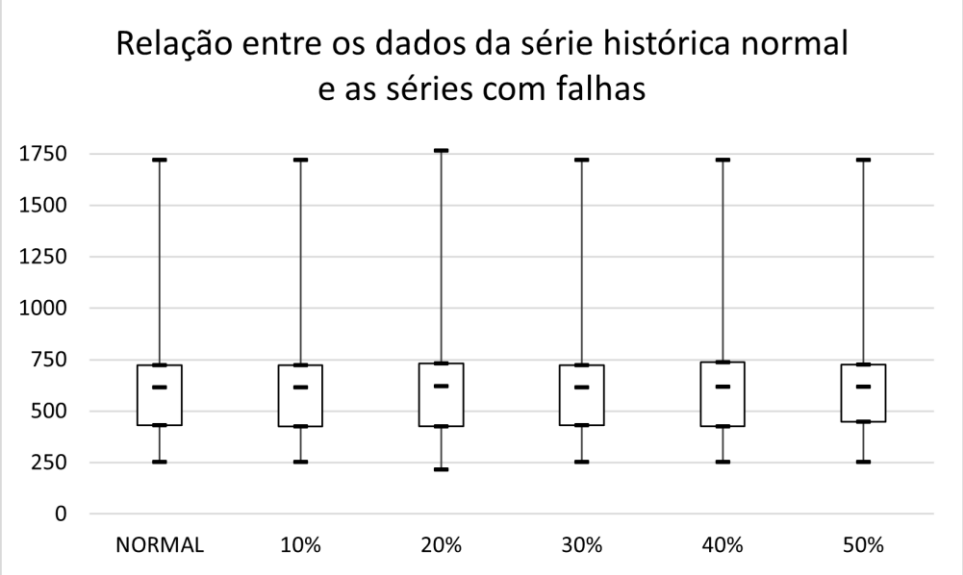
Figura 2. Distribuição acumulada para falhas na série histórica.

Org. Os autores (2023).

Falhas fictícias acima de 50% não foram utilizadas devido ao fato de que o número de falhas estaria quantitativamente elevado para que possa ser estimado. Assim, para os autores dessa pesquisa, uma possível estimação dos dados acima desse limiar não seria mais recomendável, considerando que a esse ponto a estação pluviométrica deveria ser considerada ineficiente ou inoperante.

Observando os dados através de um *boxplot* (figura 3), é possível observar, de forma semelhante, uma relação entre os dados normais do posto Piquet Carneiro e os dados de falhas fictícias, indicando ainda que o método de RLM não trouxe variações consideráveis em sua integralidade, mas indica visivelmente que apenas na faixa de falha de 20% se observa uma amplitude levemente maior que as demais, indicando que a estimação dos dados nesta faixa identificou dados menos precisos com relação aos dados normais e até mesmo em relação às outros faixas de falhas.

Figura 4. *Boxplot* com a relação entre os dados falhados e os dados normais



Org. Os autores (2023).

De forma mais detalhada, se observa no quadro 2 os dados relativos aos índices de erros relativos. O intuito deste quadro é demonstrar quais os desvios dos dados e qual a faixa de falhas tem a melhor capacidade de estimar os dados. Conforme se constata, o percentual de falha mais adequado para se utilizar a RLM é a faixa de falhas de 30%, enquanto que na faixa de 20% estão aqueles menos adequados e já eram perceptivos nas análises anteriores.

Quadro 2. Análise dos indicadores de erros para cada faixa de falhas fictícias.

FAIXA DE FALHAS	MENOR ERRO RELATIVO	MAIOR ERRO RELATIVO	MÉDIA ERRO RELATIVO	DESVIO PADRÃO ERRO RELATIVO	COEFICIENTE VARIAÇÃO ERRO RELATIVO
10%	1.6%	68.8%	20.9%	25.5%	21.9%
20%	2.1%	77.8%	17.5%	22.6%	28.6%
30%	0.0%	15.6%	3.7%	4.1%	9.4%
40%	0.3%	32.0%	5.6%	6.6%	8.3%
50%	0.1%	26.4%	5.4%	6.1%	3.9%

Org. Os autores (2023).

Comparando todas as faixas de falhas, nota-se que o modelo de RLM tem baixa capacidade de estimação com faixas mais baixas e depois tem melhora significativa nas faixas médias-altas, e posteriormente estabiliza-se nas faixas de falhas mais elevadas (até 50%).

Observando algumas pesquisas relativas aos métodos preenchimento de falhas, observa-se que o fato e que que quanto maior o percentual de falhas, menores são os erros relativos é comum em alguns métodos (ACOCK e PACHEPSKY, 2000; TEEGAVARAPU e CHANDRAMOULI, 2005).

Porém, o padrão observado nos indicadores de erros relativos pode ter características particulares para a região de pesquisa, pois os resultados podem variar de um posto para outro, pois depende de fatores como quantidade da variável, quantidade de falhas e de características espacial e temporal do posto, bem como se o mesmo está exposto a oscilações naturais do clima que afetem esse valor da variável.

Analisando problemas diversos nos trabalhos de Stock e Watson (2010), surge uma hipótese para o fato das menores faixas de falhas terem maus desempenhos nos erros relativos. Isso ocorre porque alguns modelos de regressão necessitam de dados próprios interrelacionados para estimar os próximos dados, assim, quanto mais dados estimados por modelos lineares, melhor a estimativa para dados futuros e melhor o ajuste dos dados já existentes.

Várias pesquisas indicam que outros métodos de estimação como a do vetor regional (KELLER FILHO *et al*., 2005) apresentam as melhores estimativas quando comparados outros métodos estatísticos, mas mesmo assim apresentam declínio na acurácia dos dados conforme a faixa de falhas aumenta.

Para analisar a medida de dispersão relativa, em percentagem, utilizou-se o coeficiente de variação que relaciona o desvio padrão dos erros relativos em relação à sua média. Os valores encontrados na região demonstram que os dados tiveram um índice de variabilidade de baixa a média, considerando que abaixo de 15% os dados são homogêneos e que entre 15% e 30% os dados tem uma média variabilidade (EVERITT e GRAHAN, 1991).

O padrão de desempenho dos demais indicadores de erros relativos elencados no quadro 1 também acompanha os dados do coeficiente de variação, confirmando que os melhores dados estão na faixa de 30% e que os dados com os piores desempenhos, ou seja, onde os dados tiveram mais discrepância estão na faixa de 20%.

Foi observado que, tanto de forma quantitativa quanto observando visualmente os gráficos e imagens, as estimativas demonstraram baixo desvio entre os dados estimados e os reais. Portanto, de uma forma geral, a modelagem por RLM aplicada à estação de Piquet Carneiro apresentou uma boa correlação em relação aos dados originais no quesito homogeneidade dos dados. Resultados similares utilizando métodos lineares de estimação foram obtidos por Sarkar *et al*. (2021) em West Bengal (Índia), por Luiz *et al*. (2010) e por Junqueira (2018).

Para observar melhor como as estatísticas relativas à distribuição de dados por RLM se distribuem, recomendar-se-ia uma análise de desempenho com intervalos mais curtos, como por exemplo de 5% ou 2% e/ou com falhas maiores, como por exemplo incluindo falhas acima de 50%.

Os resultados demonstram ainda um padrão comum ao semiárido brasileiro, pois modelos de estimação aplicados neste clima geralmente possuem coeficientes de variação elevados em função da alta variabilidade pluviométrica, diferentemente daqueles aplicados a regiões subtropicais que geralmente possuem coeficiente de variação mais baixos devido à forte homogeneidade pluviométrica (STUDZINSKI, 1995; NÓBREGA e SANTIAGO, 2014).

Além do padrão comum ao clima, o padrão também pode ser, em hipótese, aplicado aos padrões dos erros relativos para quaisquer regionalizações hidrológicas no semiárido utilizando-se de RLM. Porém para atestar essa hipótese faz-se necessário, sobretudo, de uma pesquisa científica mais rebuscada de pesquisas ao longo no semiárido.

Além daqueles fatores locais e regionais que podem afetar os erros relativos, outros fatores globais também podem interferir no preenchimento de falhas que é a dinâmica atmosférica. Pode-se falar por exemplo no El Niño Oscilação Sul (ENOS) e nas Anomalias na Temperatura da Superfície do Mar (ATSM).

Assim, anos anteriores ou posteriores a um dado com falha que tiveram desvios pluviométricos sujeitos ao ENOS ou ao ATSM podem interferir nos simulados daquele ano ausente, assim,

necessitando a modelagem estatística escolhida de novos ajustes nas equações ou da aplicabilidade de procedimentos computacionais de aprendizado ou treinamento e inteligência artificial.

No semiárido, a presença de falhas em dados pluviométricos pode ser facilmente ser planejada por parte do operador, pois a estação meteorológica pode ter falhas ou manutenções no período seco sem prejuízo ao fornecimento de dados anuais. É o que acontece por exemplo em alguns dados de série histórica fornecidos pela FUNCEME onde algumas falhas para algumas estações só existem em período seco.

Análise de qualidade do desempenho do preenchimento de falhas

Para estabelecer a confiabilidade e capacidade de predição da RLM na regionalização hidrológica, ou seja, testar o quanto suas previsões estão próximas dos valores reais, é necessário um conjunto de observações. Assim, para tanto, utilizou-se de métricas estatísticas de avaliação de desempenho dos dados que foram o MAE, o RMSE, o SS e o PCC.

A técnica de preenchimento de falhas por regressão teve bons desempenhos ao se calcular o MAE e a RMSE conforme se observa no quadro 3. Como a medida destes dois índices é a mesma utilizada nas viaráveis, o seu resultado pode ser empírico, a depender da amplitude dos dados utilizados pelo pesquisador. Neste caso, considerou-se que desvios acima ou abaixo de 8mm, em um universo de média pluviométrica anual de 730mm, seria considerado de excelência, representando pouco mais de 1% de desvio pluviométrico.

Quadro 3. Análise do desempenho de acurácia da estimação dos dados.

Critérios de avaliação de desempenho	**Técnica de preenchimento de falha**	**Interpretação**
	RLM	
MAE (mm)	4.91	Empírico (<8,0mm = ideal)
RMSE (mm)	7.44	Empírico (<8,0mm = ideal)
SS (adimensional)	0.22	Idealmente 1,0
PCC (adimensional)	0.30	Idealmente 1,0

Org. Os autores (2023).

Contudo, os dados relativos ao índice SS e o PCC na estação Piquet Carneiro apresentaram um desempenho ruim. Como os dados estão próximos de 0, eles indicam que há uma correlação fraca a muito fraca entre as variáveis, indicando que a probabilidade de os dados corresponderem a realidade são pouco prováveis.

Os dados de RMSE e MAE confirmam que a utilização da RLM tem como intuito evitar uma extrapolação excessiva dos dados de modo a reduzir a amplitude e o distanciamento dos dados reais. Enquanto que os dados de SS e PCC indicam que apesar dessa vantagem, estes dados não necessariamente representam valores aproximados ou utilizáveis, mas que, a RLM apenas evita que os dados estimados sejam muito sobrepujados.

Assim, observa-se que o desempenho da técnica de RLM para o preenchimento de falhas demonstrou-se não recomendado para preenchimento de falhas em dados pluviométricos após o teste de qualidade dos dados estimados, apesar da vantagem do método provocar pouco desvio entre dados estimados comparado com o dado normal ao longo das faixas de falhas.

CONSIDERAÇÕES FINAIS

Uma das maiores vantagens da aplicação do modelo de RLM é a possibilidade de gerar estimativas com maior facilidade. Esta operação estatística já está inclusa em praticamente todos os programas e ambientes de geoprocessamento. A sua aplicabilidade nesta pesquisa demonstrou-se inadequada, pois, apesar de indicar dados com um desvio muito baixo em relação aos dados normais, maximizando a verossimilhança e partindo da regionalização hidrológica, estes dados não são necessariamente próximos a realidade.

Além do mais, a RLM por sua necessidade de diversas variáveis, sofre grande influência das mesmas, assim, oculta possíveis fatores de interferência da própria região que possui falhas, como por exemplo o relevo. Conforme vão sendo adicionadas novas variáveis ao modelo, que devem ser obrigatoriamente lineares, a qualidade dos dados em regressão tende a diminuir, mesmo que a nova variável regionalizada seja importante ou muito próxima do posto de pesquisa. Isto demonstra que existe um menor controle na qualidade das estimações, bem como a possibilidade de inexatidão nas conclusões.

Os resultados dos erros esperados para a estação de Piquet Carneiro podem ser um pressuposto para um padrão no semiárido, já que em regiões de variabilidade menor como regiões subtropicais pode haver um prognóstico diferente utilizando este mesmo método, podendo ser comprovado comparando a aplicação de testes de avaliação de desempenho dos dados estimados o que inclui a análise dos erros relativos que são superiores a mesmos testes aplicados em outras regiões onde não há grande variabilidade climática.

O uso de inteligência artificial com o aprendizado computacional como redes neurais artificiais ou máquina de vetores e suporte, pode ser uma solução para aprimorar a acurácia no preenchimento de falhas com RLM ao possibilitar a introdução de um conjunto de simulações ao teste de regressão e também possibilitar a introdução de outros fatores que possibilitem melhorar o processo de regionalização como adicionar o critério do ENOS, TSM e fatores físicos e/ou meteorológicos sob a forma de camadas de entrada na arquitetura de inteligência artificial.

REFERÊNCIAS

ACOCK, M. C.; PACHEPSKY, Y. A. A. Estimating missing weather data for agricultural simulations using group method of data handling. Journal Climate Applied Meteorology and Climatology, v. 39, p. 1176–1184, 2000

ANTHES, R. A.; KUO, Y.A.; HSIE, E. Y.; LOW-NAM, S.; BETTGE, T.W. Estimation of Skill and Uncertainty in Regional Numerical Models. Quarterly Journal of the Royal Meteorological Society, v. 115, p. 763-806, 1989.

BERTONI, J. C.; TUCCI, C. E. M. Precipitação. In: TUCCI, C. E. M. (Ed.) Hidrologia: Ciência e Aplicação. Porto Alegre: UFRGS, p. 177-241, 2007.

BIER, A. A.; FERRAZ, S. E. T. Comparação de metodologias de preenchimento de falhas em dados meteorológicos para estações no sul do Brasil. Revista Brasileira de Meteorologia, v. 32, n. 2, p. 215-226, 2017.

CECÍLIO, R.A.; PRUSKI, F.F. Interpolação dos parâmetros da equação de chuvas intensas com uso do inverso de potências da distância. Revista Brasileira de Engenharia Agrícola e Ambiental, v. 7, n. 3, p. 501-504, 2003.

DEPINÉ. H.; CASTRO, N.M.R.; PINHEIRO, A.; PEDROLLO, O. O. Preenchimento de falhas de dados horários de precipitação utilizando redes neurais artificiais. Revista Brasileira de Recursos Hídricos, v. 19, n. 1, p. 51-63, 2014.

EVERITT, B.S.; GRAHAM, D. Applied multivariate data analysis. London: Edward Arnold, 1991. 354p.

FILL, H. D. Informações hidrológicas. In: BARTH, F. T.; POMPEU, C. T.; FILL, H. D.; TUCCI, C. E. M.; KELMAN, J. BRAGA JUNIOR, B. P. F. Modelos para gerenciamento de recursos hídricos. São Paulo: Nobel/ABRH, 1987. p.95-202.

JUNQUEIRA, R.; SILVA, J. A. da; OLIVEIRA, A. S. de. Comparação entre diferentes metodologias para preenchimento de falhas em dados pluviométricos. Sustentare, v. 2, n. 1, p. 198-210, 2018.

KELLER FILHO, T.; ASSAD, E. D.; LIMA, P. R. S. de R. Regiões pluviometricamente homogêneas no Brasil. Pesquisa Agropecuária Brasileira, v.40, n.4, p.311-322, 2005.

KHOSRAVI, G.; NAFARZAGEDAN, A. R.; NOHEGAR, A.; FATHIZADEH, H.; MALEKIAN, A. A. Modified distance-weighted approach for filling annual precipitation gaps: application to different climates of Iran. Theoretical and Applied Climatology, v. 119, n. 1-2, p. 33-42, 2015.

LI, M.; SHAO, Q.; ZHANG, L.; CHIEW, F.H.S. A new regionalization approach and its application to predict flow duration curve in ungauged basins. Jornal de hidrologia., v. 389, n. 1-2, p. 137-145, 2010.

MARENGO, J. A.; ALVES, L. M.; BESERRA, E.; LACERDA, F. Variabilidade e mudanças climáticas no semiárido brasileiro, In: MEDEIROS, S. de S., GHEYI, H.R., GALVÃO, C. de O., PAZ, V. P da S. (Orgs.). Recursos Hídricos em Regiões Áridas e Semiáridas. INSA, Campina Grande - PB, pp. 383- 416, 2011.

MELLO, Y. R. de; KOHLS, W.; OLIVEIRA, T. M. N. de. Uso de diferentes métodos para o preenchimento de falhas em estações pluviométricas. Boletim de Geografia, v. 35, n. 1, p. 112-121, 2017.

MILOVANOVIĆ, B.; SCHUSTER, P.; RADOVANOVIĆ, M.; RISTIĆ, V. V.; SCHNEIDER, C. Spatial and temporal variability of precipitation in Serbia for the period 1961–2010. Theoretical and Applied Climatology. v.130, n.1, p. 1-14. 2017.

MWALE, F. D.; ADELOYE, A. J.; RUSTUM, R. Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi – A self organizing map approach. Physics and Chemistry of the Earth, v. 50, p. 34-43, 2012.

NÓBREGA, R. S.; SANTIAGO, G. A. C. F. Tendência de temperatura na superfície do mar nos oceanos Atlântico e Pacífico e variabilidade de precipitação em Pernambuco. Mercator. Fortaleza, v.13, n.1, 2014. p.107-118.

OLIVEIRA, L. de.; FIOREZE, A. P.; MEDEIROS, A. M. M.; SILVA, M. A. S. Comparação de metodologias de preenchimento de falhas de séries históricas de precipitação pluvial anual. Revista Brasileira de Engenharia Agrícola e Ambiental, v. 14, p. 1186-1192, 2010.

RODRIGUES, V. C. Análise e processamento de dados hidrológicos da bacia do Rio do Carmo (MG). Relatório Técnico. Departamento de Engenharia Civil/Escola de Minas/Universidade Federal de Ouro Preto, 2002. 36p.

PAZ, A. R. Simulação Hidrológica de Rios com Grandes Planícies de Inundação. Tese de Doutorado, Programa de Pós-Graduação em Recursos Hídricos e Saneamento Ambiental, UFRGS, Porto Alegre, 258 p., 2010.

SAF, B. Assessment of the effects of discordant sites on regional flood frequency analysis. Journal of Hydrology, v. 380, n. 3-4, 2010, p. 362–375.

SARKAR, D., SARKAR, T., SAHA, S., MONDAL, P. Compiling non-parametric tests along with CA-ANN model for precipitation. Water Cycle, v.2, n.1, 2021. p. 71-84.

STOCK, J. H.; WATSON, M. W. Introduction to Econometrics. 3. ed. New York: Addison-Wesley Series in Economics, 2010.

STUDZINSKI, C. D. Um estudo da precipitação na região Sul do Brasil e sua relação com o Oceano Pacífico e Atlântico Tropical e Sul. 1995. 87 f. Dissertação (Mestrado em Meteorologia) – Instituto Nacional de Pesquisa Espacial, São José dos Campos, 1995.

TEEGAVARAPU, R. S.V.; CHANDRAMOULI, V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of Hydrology, v. 312, p. 191-206, 2005.

TUCCI, C. E. M. Regionalização de vazão. Porto Alegre: UFRGS, 256p. 2002.

VALERIANO, M. M. e ROSSETTI, D. F. 2011 - Topodata: Brazilian full coverage refinement of SRTM data. Applied Geography. v.32, n.1, p. 300-309, 2011.

ZAREKARIZI, M.; RANA, A.; MORADKHANI, H. Precipitation extremes and their relation to climatic indices in the Pacific Northwest USA. Climate Dynamics, v. 50, n. 11, p. 4519-4537, 2018.