ENTRELUGAR

DADOS DISCREPANTES OU *OUTLIER*S: AVALIAÇÃO DA QUADRA CHUVOSA DO SEMIÁRIDO DO RN, BRASIL

DISCREPANT DATA OR *OUTLIER*S: RAINY SEASON EVALUATION IN THE SEMIARID REGION OF RN, BRAZIL

DATOS DISCREPANTES O ATÍPICOS: EVALUACIÓN DE LA ESTACIÓN DE LAS LLUVIAS EN LA REGIÓN SEMIÁRIDA DE RN, BRASIL

Bruno Claytton Oliveira da Silva
Doutor em Geografia. Secretaria de Estado da Educação, da Cultura, do Esporte e do Lazer do Rio Grande do Norte (SEEC-RN).
brunoclaytton@yahoo.com.br

Ranyére Silva Nóbrega
Doutor em Meteorologia. Universidade Federal de Pernambuco (UFPE)
ranyere.nobrega@ufpe.br

RESUMO

O trabalho teve como objetivo avaliar a presença de dados discrepantes no comportamento da precipitação pluvial acumulada na quadra chuvosa (fevereiro, março, abril e maio, FMAM) do Semiárido do estado do Rio Grande do Norte (RN). Para tanto, foram empregados dados (referente ao período de 1998-2017) obtidos junto ao Banco de Dados Meteorológicos para Ensino e Pesquisa, do Instituto Nacional de Meteorologia (BDMEP-INMET), relativos às Estações Climatológicas de: Apodi-RN, Caicó-RN, Cruzeta-RN, Florânia e Macau-RN. Além disso, foram realizadas análises tanto qualitativas – a partir dos gráficos de Caixa ou *Boxplot* e de Pontos ou *Dotplots* – quanto quantitativas – por meio dos testes do Escore Z, Escore Z Modificado, Grubbs e Dixon. A partir de tal arcabouço, não foram identificados dados discrepantes em nenhuma das séries temporais e estações analisadas. Tal resultado enseja que, apesar de suas diferenças tanto metodológicas quanto em relação aos pressupostos que as sustentam, tais técnicas podem sinalizar para resultados muito próximos (ou mesmo iguais) entre si. Ainda em relação aos resultados, assegura-se que os totais anuais pluviais acumulados na estação chuvosa não apresentaram extremos significativos, sejam eles positivos ou negativos (acima/abaixo do habitual ou do excepcional). Finalmente, entende-se que estes resultados devem ser avaliados com cautela, haja vista que tal "estabilidade" temporo-espacial não necessariamente representa um aspecto favorável, por exemplo, para a gestão das águas – sobretudo, quanto ao abastecimento público – do Semiárido do RN, haja vista que, para este caso, acumulados discrepantes positivos são desejáveis.

Palavras–chave: Dados Discrepantes; Precipitação Pluvial; Quadra Chuvosa; Semiárido; Rio Grande do Norte.

ABSTRACT

The objective of this study was to evaluate the presence of discrepant data in the behavior of the accumulated rainfall in the rainy season (February, March, April, and May, FMAM) of the Semi-arid region of the state of Rio Grande do Norte (RN). For this, data were employed (referring to the period 1998-2017) obtained from the Meteorological Database for Teaching and Research, of the National Institute of Meteorology (BDMEP-INMET), relating to the Climatological Stations of: Apodi-RN, Caicó-RN, Cruzeta-RN, Florânia and Macau-RN. In addition, both qualitative - from the Boxplot and Dotplots - and quantitative - through the Z-score, Modified Z-score, Grubbs, and Dixon tests - analyses were carried out. Based on this framework, no discrepant data were identified in any of the time series and stations analyzed. This result implies that, despite their methodological differences as well as in relation to the assumptions that support them, these techniques can signal very close (or even equal) results between them. Still in relation to the results, it is assured that the accumulated annual rainfall totals in the rainy season did not present significant extremes, whether positive or negative (above/below the usual or exceptional). Finally, it is understood that these results should be evaluated with caution, given that such temporal-spatial "stability" does not necessarily represent a favorable aspect, for example, for the management of water - especially for public supply - in the Semi-arid region of RN, given that, for this case, positive discrepant accumulations are desirable.

Keywords: Discrepant Data; Precipitation; Rainy Season; Semi-arid; Rio Grande do Norte.

RESUMEN

El objetivo de este estudio fue evaluar la presencia de datos discrepantes en el comportamiento de la precipitación acumulada en la temporada de lluvias (febrero, marzo, abril y mayo, FMAM) de la región semiárida del estado de Rio Grande do Norte (RN). Para ello, se han utilizado datos (referidos al periodo 1998-2017) obtenidos de la Base de Datos Meteorológicos para la Docencia y la Investigación, del Instituto Nacional de Meteorología (BDMEP-INMET), relativos a las Estaciones Climatológicas de: Apodi-RN, Caicó-RN, Cruzeta-RN, Florânia y Macau-RN. Además, se realizaron análisis tanto cualitativos -a partir de los Boxplot y Dotplots- como cuantitativos -a través de las pruebas Z-score, Z-score modificado, Grubbs y Dixon-. A partir de este marco, no se identificaron datos discrepantes en ninguna de las series temporales y estaciones analizadas. Este resultado implica que, a pesar de sus diferencias tanto metodológicas como en relación con los supuestos que las sustentan, dichas técnicas pueden señalar resultados muy cercanos (o incluso iguales) entre sí. También en relación con los resultados, se asegura que los totales de precipitación anual acumulada en la temporada de lluvias no mostraron extremos significativos, ya sean positivos o negativos (por encima/por debajo de lo habitual o excepcional). Por último, se entiende que estos

resultados deben ser evaluados con cautela, dado que dicha "estabilidad" temporal-espacial no representa necesariamente un aspecto favorable, por ejemplo, para la gestión del agua -especialmente en lo que respecta al abastecimiento público- en la región semiárida de RN, dado que, para este caso, son deseables las acumulaciones discrepantes positivas.

Palabras clave: Datos discrepantes; Precipitación; Temporada de lluvias; Semiárido; Rio Grande do Norte.

INTRODUÇÃO

A análise de dados discrepantes, objetivamente, visa avaliar a presença/ausência de dados aberrantes, atípicos ou *outliers* (termo usado em inglês) em uma ou mais séries estatísticas. Logo, tal análise figura como importante, entre outros aspectos, no processo de tratamento estatístico daqueles (SILVA, 2019).

Sobre a origem e possíveis explicações para a presença de dados discrepantes (*outliers*) em um conjunto de dados, Andriotti (2005, p. 24) menciona que:

> *Outliers* são tão diferentes dos demais valores disponíveis para estudo que se pode suspeitar que sejam oriundos de alguma falha ou mesmo anormalidade na aplicação do teste aplicado, ou ainda estar-se na presença de uma observação que não pertence ao grupo de estudo [...] Dentre as várias possíveis fontes deste tipo de dado se pode citar a presença de erros analíticos, contaminação, erros de digitação e/ou transição de resultados, e erros de interpretação, como classificar erroneamente determinado grupo, incluindo seus valores em outro (ANDRIOTTI, 2005, p. 24).

Como citado, são várias as possíveis explicações para a presença de *outliers* em uma série. Todavia, deve-se destacar que o processo de eliminação ou manutenção desse(s) nas distribuições, deve ser acompanhado de critérios técnicos – específicos a dada área –, e/ou através da aplicação de técnicas desenvolvidas para tal (SILVA, 2021).

Do contrário, como tais dados influenciam, inclusive, nos valores médios, nas medidas de dispersão e nas correlações com outras variáveis, sua manutenção ou remoção poderá produzir conclusões que não dizem respeito ao conjunto dos dados (ANDRIOTTI, 2005).

Assim como ocorre em outras apreciações, a análise de dados discrepantes pode ser realizada a partir de recursos gráficos (análises qualitativas), como também por meio do emprego de formulações estatísticas (técnicas quantitativas).

Fundamentalmente, na perspectiva de análise qualitativa dos dados, a principal representação gráfica utilizada para tal é a proposta por Tukey (1977), que utiliza gráficos de Caixa ou, em inglês, *Boxplot*. Secundariamente, utilizam-se, também, os gráficos de Pontos ou, em inglês, *Dotplot* (SILVA, 2021).

Por outro lado, em avaliações que primam pelo rigor quantitativo, existe uma série de técnicas para tal, dentre elas: o teste do Desvio Padrão (DP) ou *Standard Deviation* (*SD*), Escore Z (*Z-Score*), e, Escore Z Modificado (*Modified Z-Score*); além dos testes de Grubbs e Dixon (ANDRIOTTI, 2005; SEO, 2006).

Notadamente, a avaliação do comportamento das precipitações pluviométricas no Semiárido brasileiro, para fins de caracterização (quantitativa), tem sido historicamente objeto de vários estudos, em particular, ligados ao período chuvoso da citada região geográfica, entre eles: Silva (2021), Carvalho (2020), Tavares, Arruda e Silva (2019), Ramalho e Guerra (2018) e Souza, Nogueira e Nogueira (2017).

Tal fato possui várias justificativas, dentre elas (MARENGO, 2008):

- além da elevada e comum variabilidade espaço-temporal das precipitações pluviais na região, há ocorrência frequente de ‘veranicos’ durante a estação chuvosa;
- o mencionado período do ano se configura como aquele em que ocorre o maior cômputo e, portanto, acúmulo de precipitação pluvial na região;
- a estação possui grande relevância para recarga dos recursos hídricos (principalmente os superficiais) e, consequentemente, para o abastecimento e dinamismo das atividades econômicas da área;
- o impacto das Secas Meteorológicas, Hidrológicas e Agrícolas sobre as Secas Sociais, acima de tudo, sobre parcela significativa da população residente na área que dependente, quase que exclusivamente, de atividades ligadas à agricultura de sequeiro e pecuária rudimentar para sua subsistência;
- recentemente, devido as ameaças das mudanças climáticas globais e seus diversos impactos previstos; e

- em função das lacunas, ainda existentes, a despeito das técnicas de análise empregadas em tais trabalhos.

Diante do exposto, objetivou-se avaliar a presença de dados discrepantes quanto ao comportamento da precipitação pluvial acumulada na quadra chuvosa (fevereiro, março, abril e maio, FMAM) do Semiárido do estado do Rio Grande do Norte (RN).

Área de Estudo

O estado do RN, unidade administrativa da federação em estudo, possui altitude máxima de 831m e 52.810,70km²; o que corresponde a 0,62% do território nacional. Seus pontos extremos/divisas, e respectivas coordenadas geográficas, são: ao norte, 04°49'53"S e 37°15'11"W; ao sul, 06°58'57"S e 36°43'01"W; a leste, 06°29'18"S e 35°58'03"W; e a oeste, 06°23'23"S e 38°36'12"W (IDEMA, 2015).

O RN se destaca como sendo o estado que possui o segundo maior percentual de municípios (147 de 167, 88%) entre aqueles que compõe a região Semiárida brasileira (SILVA, 2021).

Mais especificamente, o recorte espacial deste trabalho é representado pelos municípios de: Apodi-RN, Caicó-RN, Cruzeta-RN, Florânia-RN e Macau-RN.

A escolha dos municípios supraditos está relacionada à presença, em seus territórios, de postos pluviométricos ou Estações Climatológicas pertencentes ao Instituto Nacional de Meteorologia (INMET). Por conseguinte, dada a relevância da entidade, tais séries temporais apresentam elevado grau de confiabilidade no que tange aos procedimentos de observação, crítica, tabulação, apuração e registro dos dados.

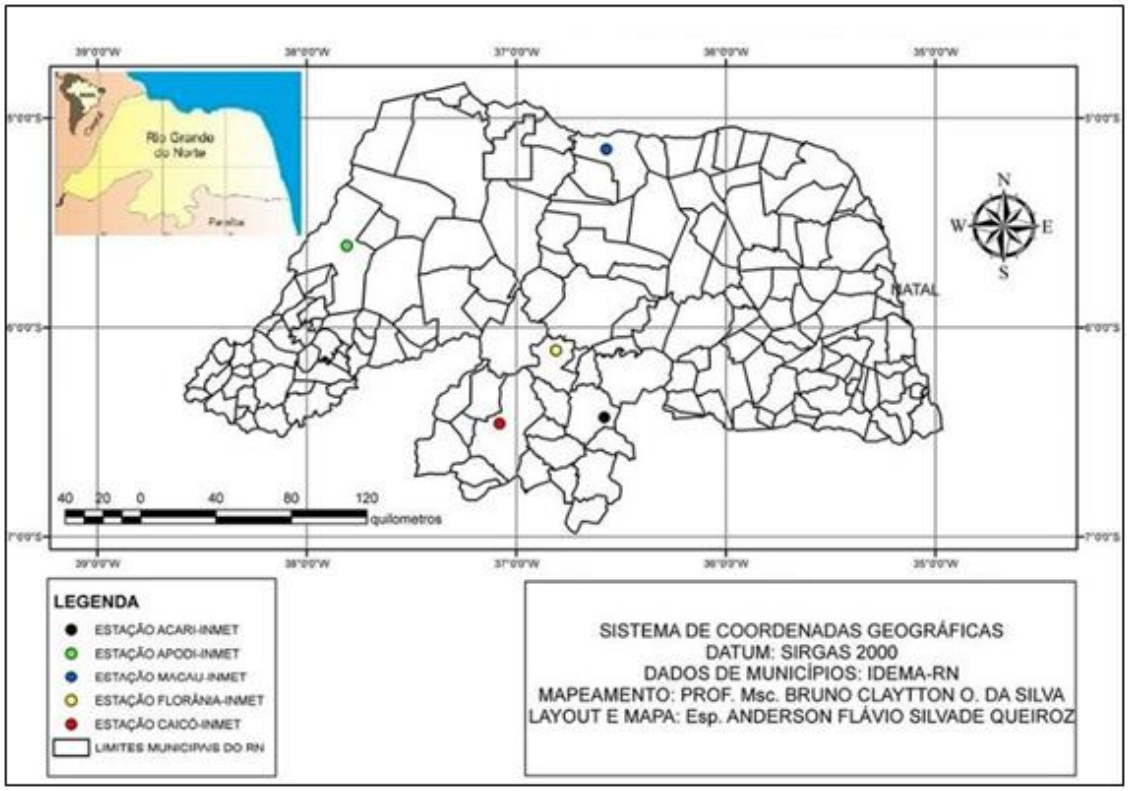
Não obstante, destaca-se que foram tabuladas e processadas cinco séries históricas, com recorte temporal de 1998 a 2017, representadas pelas Estações Climatológicas de: Apodi, Caicó, Cruzeta, Florânia e Macau.

As cinco estações analisadas possuem as seguintes denominações, municípios, numerações-padrão da Organização Mundial de Meteorologia (OMM), coordenadas geográficas, altitudes e situação quanto à operação, respectivamente (BDMEP, 2015):

- Estação Apodi (em Apodi, OMM 82590, operante), Lat. -5,61° e Log. -37,81° e 150,00m;
- Estação Seridó (em Caicó, OMM 82690, operante), Lat. -6,46° e Log. -37,08° e 169,85m;
- Estação Cruzeta (em Acari, OMM 82693, operante), Lat. -6,43° e Log. -36,58° e 226,46m;
- Estação Florânia (em Florânia, OMM 82691, operante), Lat. -6,11° e Log. -36,81° e 324,45m;
- Estação Macau (em Macau, OMM 82594, operante), Lat. -5,15° e Log. -36,57° e 32,00m.

Na figura 1, tem-se a localização das estações utilizadas no trabalho:

Figura 1 – Mapa de Localização das Estações Climatológicas do Semiárido Potiguar



Fonte: Silva (2021).

METODOLOGIA

Proposta de Tukey (1977): Gráficos de Caixa ou *Boxplot*

A proposta de Tukey (1977), mais conhecida como gráficos de Caixa ou *Boxplot*, é empregada para avaliação qualitativa, entre outros, da presença/ausência de dados discrepantes em séries diversas.

Discorrendo sobre o processo de construção, estrutura e informações inerentes a tal representação gráfica, Magalhães e Lima (2013, p. 20) afirmaram que:

> Para construção do *box-plot*, definimos um retângulo ("caixa") em que a aresta inferior coincide com o primeiro quartil e a superior, com o terceiro quartil. A mediana é representada por um traço no interior do retângulo. Segmentos de reta, denominados de bigodes por alguns autores, são incluídos no *box-plot*, partindo dos primeiro e terceiro quartis [...] são limitados pelos valores mínimo e máximo [...].

A partir da última citação, concluir-se-á que a conjunção da "Caixa" mais os seus "bigodes", apresentarão o "conjunto dos dados", com exceção feita àqueles que serão denominados de Potencialmente Atípicos, Aberrantes, Discrepantes ou *Outliers*. Esses, no *Boxplot*, serão representados, geralmente, por um asterisco (*), sendo posicionados para além dos "bigodes.

A construção do *Boxplot*, segundo a técnica de Tukey (1977), está relacionada a observação da diferença entre os valores do terceiro e do primeiro Quartis dos dados; denominada de Intervalo Interquartil ou Interquartílico ou Amplitude Interquartil ou Interquartílica. O mesmo é calculado por (PORTAL ACTION, 2017a):

$$IQ = Q_3 - Q_1 \qquad (1)$$

Onde: IQ = Intervalo Interquartil ou Interquartílico; Q_3 = Terceiro Quartil; Q_1 = Primeiro Quartil.

O IQ representa a concentração de 50% das observações centrais do *Boxplot*, expressando, assim, a dispersão dos dados observados no intervalo. Todavia, o IQ,

simplesmente, não é útil para se observar a presença de *outliers*, mas figura como um componente fundamental para sua identificação.

Tal fato justifica-se pela presença do IQ na formação empregada para definição dos pontos de corte, ou seja, para delimitação do intervalo no qual os dados estariam dentro de um padrão esperado. Logo, tais limites são definidos por (MAGALHÃES E LIMA, 2013, p. 20; PORTAL ACTION, 2017a):

$$LI = [Q_1 - 1{,}5 * IQ] \qquad (2)$$

Onde: LI = Limite Inferior; Q_1 = Primeiro Quartil; IQ = Intervalo Interquartil.

$$LI = [Q_1 + 1{,}5 * IQ] \qquad (3)$$

Onde: LS = Limite Superior; Q_3 = Terceiro Quartil; IQ = Intervalo Interquartil.

Após o cálculo dos limites acima apresentados, deve-se observar se existem dados acima do LS ou abaixo do LI. Caso exista(m), esse(s) será(ão) considerado(s) discrepante(s), atípico(s) ou *outlier(s)* (MAGALHÃES E LIMA, 2013).

Gráficos de Pontos ou *Dotplots*

Os gráficos de Pontos ou *Dotplots*, destacam-se como uma outra representação gráfica que permite a avaliação, entre outras funções, da distribuição de uma série de dados com vistas a análise de *outliers*.

Os *Dotplots* representam cada observação obtida em uma escala horizontal, permitindo visualizar a série dos dados ao longo deste eixo. Acrescenta-se que, no eixo horizontal, divide-se a escala dos valores em intervalos, sendo marcado um ponto por observação (PORTAL ACTION, 2017b).

Numa outra perspectiva, pode-se entender que os *Dotplots* são diagramas pontuais que permitem a visualização horizontal de como as observações da variável se distribuem na reta (SILVESTRE, SANT'ANNA NETO E FLORES, 2013).

Os Testes do Desvio Padrão, do Escore Z e do Escore Z Modificado

Um dos testes utilizados para análise de *outliers* é o teste do Escore Z Modificado. Seu emprego se justifica, pois, o mesmo utiliza estimadores robustos para tal propósito, como a Mediana.

Como ressalta Andriotti (2005, p. 25):

> Este teste tem sido usado de forma mais extensa que o teste que considera como *outlier* simplesmente os valores que superam a soma da média aritmética com três desvios padrão, ou a média menos três desvios padrão, pois que tanto a média como o desvio padrão são, já, afetados pela presença de *outlier*.

A partir da citação, pode-se perceber que o emprego de tal técnica garantirá que os limites definidos para se identificar os *outliers* não sejam influenciados por eles mesmos; como pode ocorrer quando se emprega tanto o teste do Desvio Padrão (DP), como também o teste do Escore Z.

O teste do DP é definido a partir da classificação descrita abaixo (SEO, 2006, p. 9):

$$2\ SD: \bar{x} \pm 2\ SD \qquad (4)$$

$$2\ SD: \bar{x} \pm 2\ SD \qquad (5)$$

Onde: $\overline{x}$ = Média Aritmética amostral; *SD* = Desvio Padrão.

Já o teste Escore Z considera a Média Aritmética e o Desvio Padrão, sendo definido por (FETTERMANN, 2015, p. 118; SEO, 2006, p. 10):

$$z_i = \frac{x_i - \bar{x}}{s}, \text{ Quando } X_i \sim N(\mu, \sigma^2) \qquad (6)$$

Onde: z_i = valor padronizado (*standardized value*); $x_i - \bar{x}$ = desvios em relação à Média Aritmética; s = Desvio Padrão (ou SD, em inglês).

O pressuposto para utilização do teste do Escore Z é o seguinte: se "x_i" for uma variável aleatória com distribuição Normal, N (μ, σ^2), "z_i" segue uma distribuição Normal Padrão, N (0,1). Logo, os Escores Z que superarem ±3 SD da Média Aritmética poderão ser considerados *outliers*.

Diante das limitações apresentadas para o teste do *Escore* Z, surgirá a necessidade de aplicação do teste do *Escore* Z Modificado como um instrumento que objetiva superar tais limitações.

Na literatura são vários os autores que apontam o teste do Escore Z Modificado como aquele que produz melhor desempenho para detecção de *outliers* em séries diversas. Logo, seus resultados possuem ampla aceitação na comunidade científica, tendo sido mencionado nos trabalhos de: Shiffler (1988), Crosby (1994), Ben-Gal (2005), Seo (2006) e Fettermann (2015).

Fettermann *et al*. (2015, p. 118) definem o Escore Z Modificado em duas etapas:

1ª etapa: $$MAD = \text{median}\{|x_i - \bar{x}_1|\} \quad (7)$$

Onde: MAD = Média Aritmética dos desvios (absolutos); *Median* = Mediana; $|x_i - \bar{x}_1|$ = módulo dos desvios.

2ª etapa: $$M_i = \frac{0,6745}{MAD}(xi - \bar{x}_1) \quad (8)$$

Onde: M_i = Escore Z Modificado; 0,6745 = constante; MAD = Média Aritmética dos desvios; $|x_i - \bar{x}_1|$ = módulo dos desvios.

Teste de Grubbs (1969)

O teste de Grubbs (1969), também conhecido como Teste Residual Normalizado Máximo, ou, em inglês, *Maximum Normed Residual Test*, é mais um teste aplicado à detecção de *outlier(s)* (NIST/SEMATECH, 2012).

Para o correto emprego da técnica de Grubbs (1969), deverão ser procedidas as seguintes etapas (ANDRIOTTI, 2005, p. 26):

1. Converter os dados reais em Logaritmos Naturais ou Neperianos (LN);
2. Calcular a Média Aritmética e o Desvio Padrão dos dados em LN;
3. Dispor os dados convertidos em Rol (ordem crescente);
4. Para suspeita de *outlier*, relacionado ao menor valor (em LN), calcula-se como (GRUBBS, 1969, p. 5):

$$\tau_G = \frac{\left[\bar{x} - x_1\right]}{s} \tag{9}$$

Onde: τ_G = valor crítico (Observado/Calculado); $\bar{x}$ = Média Aritmética dos dados convertidos em LN; x_1 = primeiro dado da série em rol (crescente); s = Desvio Padrão amostral.

5. Para suspeita de *Outlier* relacionado ao maior valor (em LN), calcula-se (GRUBBS, 1969, p. 4):

$$\tau_G = \frac{\left[x_n - \bar{x}\right]}{s} \tag{10}$$

Onde: x_n = último dado da série em rol (crescente).

6. definir um certo nível de significância (α), observar o "n" em questão e verificar o valor crítico tabelado ("τG" tabelado); e

7. comparar o valor crítico calculado com o tabelado. Se o valor crítico calculado/observado for maior que o tabelado, rejeita-se a hipótese nula (H0) e assume-se que aquele dado é realmente um *outlier*.

Teste de Dixon (1950)

O teste de Dixon (1950) é mais um teste tradicional empregado para detecção de *outlier(s)*. Tal teste destaca-se, em relação àqueles até então apresentados, por não demandar o conhecimento do Desvio Padrão Amostral (BORGES, 2006).

Segundo Andriotti (2005, p.28), "o teste de Dixon (1950) é usado mais comumente na detecção de pequenas quantidades de *outliers*, e recomendado quando o número de observações está entre 3 e 25 [...]". Todavia, segundo Hawkins (1980, p. 35), essa estatística "*assume normality* [...]". Além disso, o mesmo autor (1980, p. 41) complementa:

> Fica assim claro que, a menos que a distribuição normal se aproxime muito bem da distribuição real nos rabos extremos, as conclusões tiradas do uso da aproximação normal podem estar erradas em grande parte de forma quase arbitrária. Assim, as aproximações normais devem ser usadas com extrema cautela [...]. (HAWKINS, 1980, p. 35, tradução nossa).

A mesma recomendação – quanto a normalidade dos dados – é feita por Andriotti (2003, p.29), que afirma que os resultados do teste são válidos "[...] para conjuntos de dados que se ajustem à distribuição normal".

A partir das últimas citações pode-se perceber que o emprego do teste de Dixon (1950) está condicionado à normalidade dos dados ou, ao menos, um comportamento desses que seja aproximado a essa distribuição. Por consequência, como mencionou Hawkins (1980), deve-se ter bastante cautela ao utilizá-la.

Para se operar com o teste de Dixon (1950) devem ser procedidas as seguintes etapas (DIXON, 1950; ANDRIOTTI, 2005; BORGES, 2006):

1. Dispor os dados em Rol (ordem crescente) – $x_1 < x_2 < x_3 < ... < x_n$;
2. Supor que o menor valor (x_1) ou o maior valor (x_n) são suspeitos de serem *outliers*;
3. De acordo com o valor de "n", para suspeita de *outlier* relacionado ao menor valor, calcular:

Quadro 1 – Equações para Obtenção do Valor Crítico Mínimo do Teste de Dixon (1950)

N	Razão	Se x_1 é suspeito	Numeração da Formulação
$3 \leq n \leq 7$	τ_{D10}	$(\mathbf{x_2 - x_1})/(\mathbf{x_n - x_1})$	(11)
$8 \leq n \leq 10$	τ_{D11}	$(\mathbf{x_n - x_1})/(\mathbf{x_{n-1} - x_1})$	(12)
$11 \leq n \leq 13$	τ_{D21}	$(\mathbf{x_3 - x_1})/(\mathbf{x_{n-1} - x_1})$	(13)
$14 \leq n \leq 25$	τ_{D22}	$(x_3 - x_1) / (x_{n-2} - x_1)$	(14)

Fonte: Borges (2006).

4. De acordo com o valor de "n", para suspeita de *outlier* relacionado ao maior valor, calcular:

Quadro 2 – Equações para Obtenção do Valor Crítico Máximo do Teste de Dixon (1950):

n	Razão	Se x_n é suspeito	Numeração da Formulação
$3 \leq n \leq 7$	τ_{D10}	$(\mathbf{x_n - x_{n-1}})/(\mathbf{x_n - x_1})$	(15)
$8 \leq n \leq 10$	τ_{D11}	$(\mathbf{x_n - x_{n-1}})/(\mathbf{x_n - x_2})$	(16)
$11 \leq n \leq 13$	τ_{D21}	$(\mathbf{x_n - x_{n-2}})/(\mathbf{x_n - x_2})$	(17)
$14 \leq n \leq 25$	τ_{D22}	$(\mathbf{x_n - x_{n-2}})/(\mathbf{x_n - x_3})$	(18)

Fonte: Borges (2006).

5. Definir um dado nível de significância (α), observar o "n" em questão e verificar o valor crítico tabelado; e

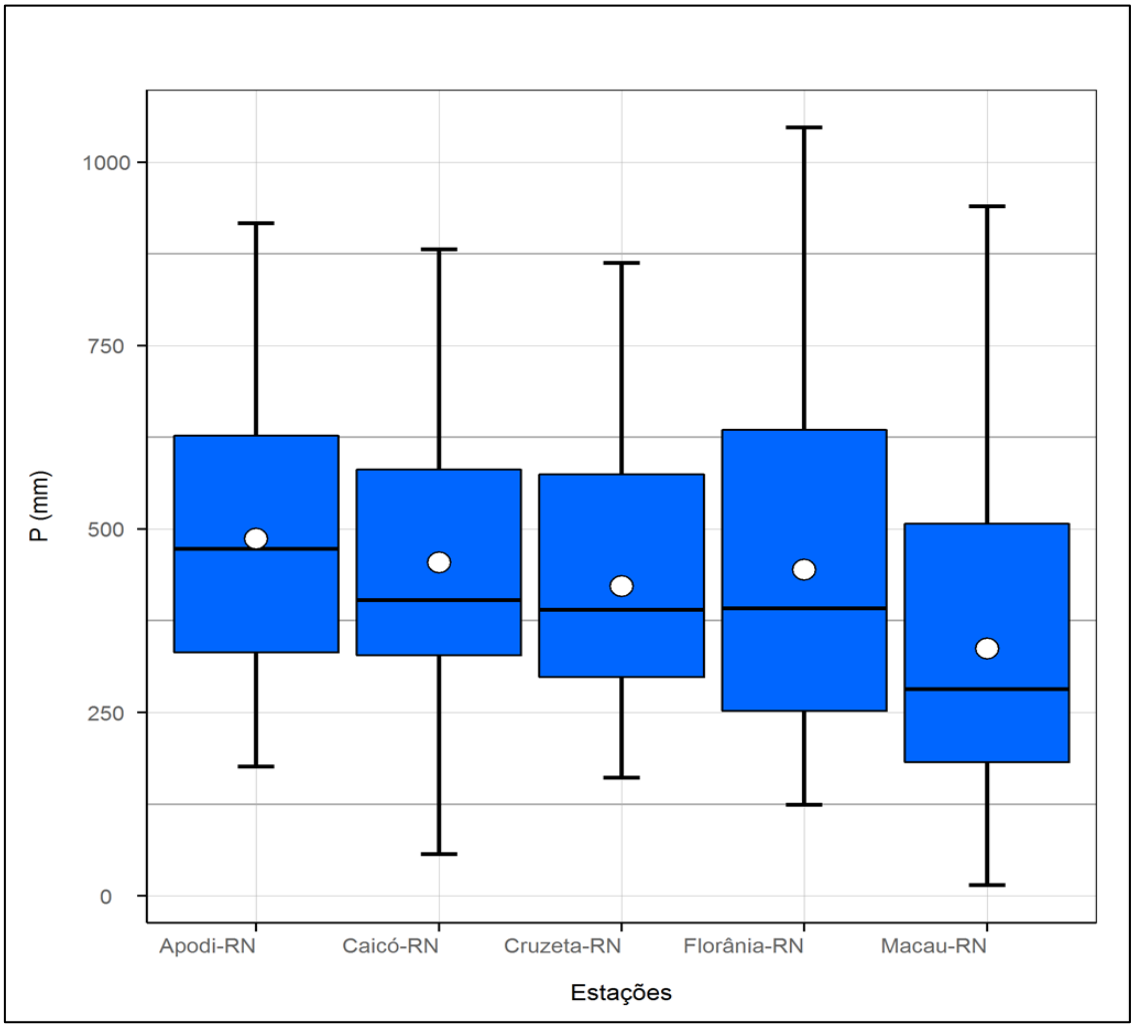
6. Comparar o valor crítico calculado com o tabelado. Se o valor crítico do primeiro for maior que o do segundo, rejeita-se a hipótese nula (H0), e assume-se que aquele dado é realmente um *outlier*.

RESULTADOS E DISCUSSÕES

Como já mencionado, foram realizadas no trabalho análises tanto qualitativas (através dos gráficos de Caixa ou *Boxplot* e de Pontos ou *Dotplots*), quanto quantitativas (testes do Escore Z, Escore Z Modificado, Grubbs e Dixon) para avaliar a presença de dados discrepantes no comportamento da precipitação pluvial acumulada na quadra chuvosa (FMAM) do Semiárido do estado do RN.

Abaixo, segue a primeira supradita representação gráfica:

Figura 2 – *Boxplots* da Precipitação Acumulada na Quadra Chuvosa (FMAM) para as Estações Analisadas (1998-2017)



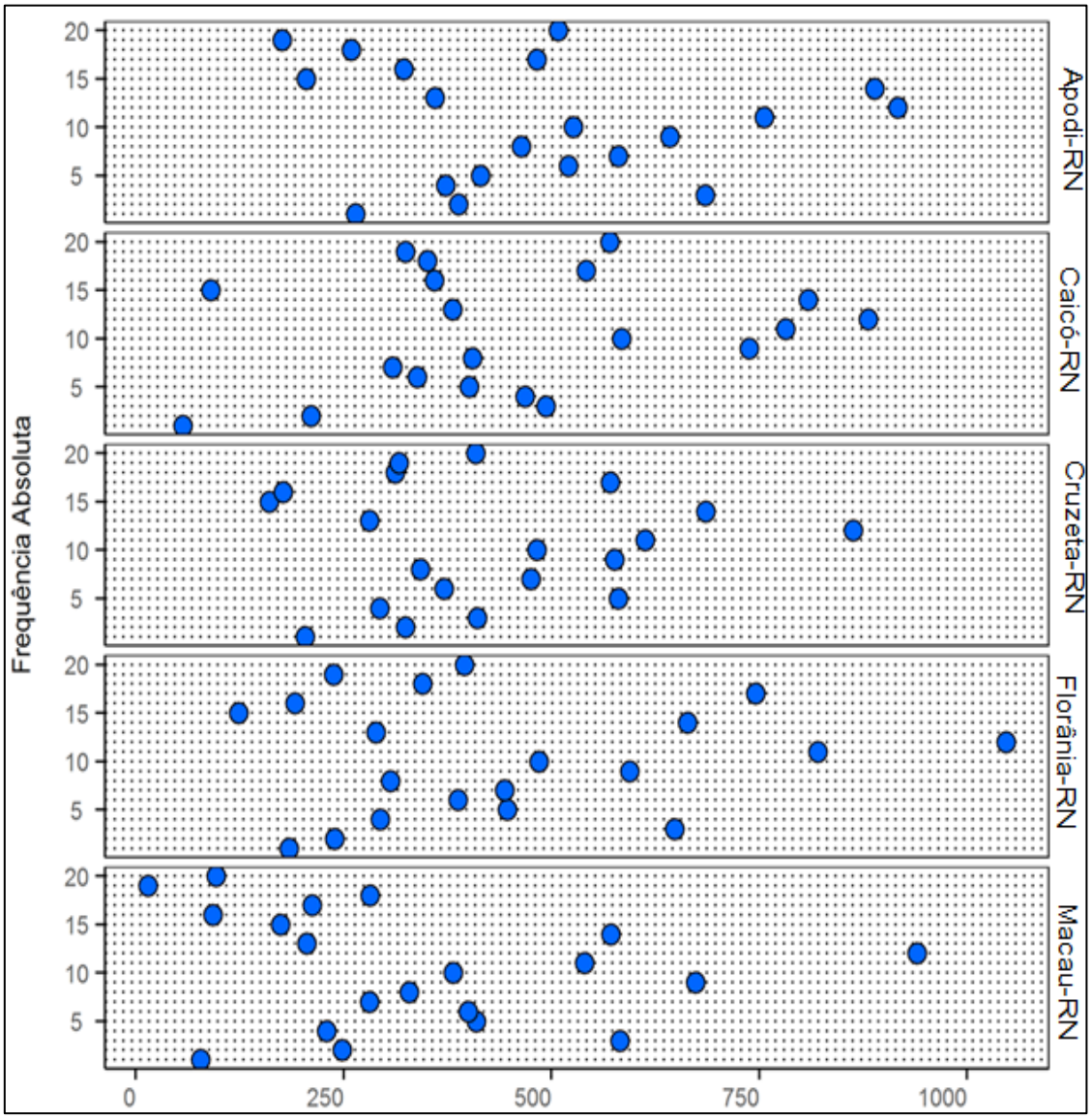
Fonte: Próprio autor, 2021.

Para fins de identificação de *outliers* no gráfico de Caixas acima exposto, deve-se observar a presença de asteriscos (*), antes ou após os seus "bigodes" (MAGALHÃES E LIMA, 2013, p. 25).

A partir do exposto, e da verificação indicada, conclui-se que, segundo o gráfico de Caixas, não há dados discrepantes em nenhuma das séries avaliadas, haja vista que nelas não se observou asteriscos.

O segundo recurso gráfico usado para detecção de dados discrepantes na quadra chuvosa do Semiárido potiguar foi o Diagrama de Pontos ou *Dotplot*; expresso abaixo:

Figura 3 – *Dotplots* da Precipitação Acumulada na Quadra Chuvosa (FMAM, 1998-2017) para as Estações Avaliadas



Fonte: Próprio autor, 2021.

De antemão, para leitura devida do recurso acima, esclarecesse que o gráfico apresenta cada uma das observações da precipitação acumulada anual (mm) – eixo horizontal – referente as estações/séries em análise; permitindo, assim, a visualização de como os dados estão organizados e dispersos.

Especialmente, para fins de análise de *outliers*, deve-se observar no *Dotplot* a dispersão dos pontos no eixo horizontal do gráfico, sobretudo os pontos isolados dos demais (PORTAL ACTION, 2017b).

Em face do exposto, e do comportamento bastante disperso dos pontos em cada estação, não se acredita ser possível (para tais dados) aceitar ou refutar, com convicção, a presença de *outliers* nas séries temporais avaliadas. Ou seja, o resultado apresentado mostra-se inconclusivo.

De modo a excluir qualquer dúvida sobre a presença de dados discrepantes, aplicou-se, também, técnicas quantitativas para tal fim, sendo as duas primeiras técnicas empregadas: os testes do Escore Z e do Escore Z Modificado.

A síntese dos seus resultados encontra-se no quadro abaixo:

Quadro 3 – Síntese dos Resultados dos Testes do Escore Z e do Escore Z Modificado.

ESTAÇÕES	**Z_i (Mín.)**	**Z_i (Máx.)**	**$\|Z_{i*}\|$ (Máx.)**	**STATUS**
Apodi-RN	-1,5	2,0	1,8	Ausentes
Caicó-RN	-1,8	1,9	1,9	Ausentes
Cruzeta-RN	-1,4	2,4	2,2	Ausentes
Florânia-RN	-1,3	2,5	2,4	Ausentes
Macau-RN	-1,4	2,6	2,6	Ausentes

Fonte: Próprio autor, 2021.

Baseando-se na fundamentação teórica, relativa as técnicas supramencionadas anteriormente, pode-se afirmar que em nenhuma das séries temporais avaliadas verificou-se dados atípicos; segundo as diretrizes dos testes do Escore Z e do Escore Z Modificado.

Tal conclusão está balizada nos intervalos tipificados para definição de dado discrepante de cada um dos testes, a saber: 1. Segundo o teste do Escore Z, $z_i < -3,0$ ou $z_i > 3,0$; 2. A partir do teste do Escore Z Modificado, $|z_{i*}| > 3,5$ (Shiffler, 1988; Iglewicz e Hoaglin, 1993; Crosby, 1994; Ben-Gal, 2005; Seo, 2006; Fettermann, 2015).

A terceira técnica utilizada para avaliação da presença de dados discrepantes nas séries estudadas foi o teste de Grubbs (1969). O resumo dos seus resultados é apresentado a seguir:

Quadro 4 – Síntese dos Resultados do Teste de Grubbs (1969)

Estações	**Precipitação Média Acumulada na Quadra Chuvosa (MM)**	**Z_{Grubbs}**	***Status***
Apodi-RN	486,7	1,02	Ausentes
Caicó-RN	454,7	0,45	Ausentes
Cruzeta-RN	422,4	0,12	Ausentes
Florânia-RN	444,5	0,27	Ausentes
Macau-RN	337,1	1,63	Ausentes

Fonte: Próprio autor, 2021.

Como já mencionado, as séries temporais avaliadas possuem o mesmo "n" (20), ou seja, são balanceadas. Portanto, para "α" igual a 0,05 (5,0%) se verificará que seu Z crítico (Z_c) é igual a 2,557. Logo, como a técnica sugere que a hipótese nula do teste (H_0) deverá ser rejeitada somente se $Z > Z_c$, concluir-se-á que não há *outliers* em nenhum dos conjuntos avaliados.

Finalmente, a quarta técnica quantitativa usada para identificação de *outliers* foi o teste de Dixon (1950). Seus resultados seguem abaixo:

Tabela 1 – Síntese dos Resultados do Teste de Dixon

ESTAÇÕES	**P (mm)**		**ESTATÍSTICA DO TESTE DE DIXON**		***STATUS***
Apodi-RN	x_1	176,0	TD	0,14	Ausentes
	x_{20}	916,6		0,24	Ausentes
Caicó-RN	x_1	56,7	TD	0,21	Ausentes
	x_{20}	881,0		0,15	Ausentes
Cruzeta-RN	x_1	160,7	TD	0,09	Ausentes
	x_{20}	862,9		0,38	Ausentes
Florânia-RN	x_1	123,9	TD	0,11	Ausentes
	x_{20}	1047,1		0,35	Ausentes
Macau-RN	x_1	14,5	TD	0,14	Ausentes
	x_{20}	939,8		0,42	Ausentes

Fonte: Próprio autor, 2021.

Inicialmente, vale destacar que os valores de "x_1" e "x_{20}", apresentados na tabela acima, referem-se, respectivamente, aos mínimos e máximos observados em cada uma das séries estudadas. Ademais, ressalta-se que as bases para avaliação do teste de Dixon foram: "n" = 20, "α" = 0,05 (5,0%) e, assim, $T_{Dtabelado} = 0,450$.

Assim, os pressupostos do teste apontam que se deverá rejeitar H_0 apenas nos casos em que for observado $T_D > T_{Dtabelado}$ (DIXON, 1950; ANDRIOTTI, 2005;

BORGES, 2006). Portanto, como em nenhuma estação isso ocorreu, conclui-se que não há *outliers* em suas séries.

CONCLUSÕES

A partir das análises qualitativas realizadas – a partir dos recursos gráficos – e do emprego de técnicas quantitativas, ambas, com vistas à detecção de dados discrepantes da precipitação pluvial acumulada na quadra chuvosa do Semiárido potiguar, não se identificou nenhum dado atípico para o recorte temporal (1998-2017) e estações analisadas.

Tal resultado enseja que, apesar de suas diferenças tanto metodológicas quanto em relação aos pressupostos que as sustentam, tais técnicas podem sinalizar para resultados muito próximos (ou mesmo iguais) entre si.

Ainda por meio dos resultados, assegura-se que os totais anuais acumulados na estação chuvosa (FMAM) não apresentaram extremos significativos, sejam eles positivos ou negativos.

Finalmente, entende-se que estes resultados devem ser avaliados com cautela, haja vista que tal "estabilidade" temporo-espacial não necessariamente representa um aspecto favorável, por exemplo, para a gestão das águas – sobretudo, quanto ao abastecimento público – do Semiárido do RN, haja vista que, para este caso, acumulados discrepantes positivos são desejáveis.

REFERÊNCIAS

ANDRIOTTI, J. L. S. **Técnicas Estatísticas Aplicáveis a Tratamento de Informações Oriundas de Procedimentos Laboratoriais**. Porto Alegre: CPRM, 2005. Disponível em: <http://rigeo.cprm.gov.br/xmlui/bitstream/handle/doc/451/Andriotti_Tecnicas_estatisticas.pdf?sequence=1>. Acesso em: 29 abr. 2017.

BEN-GAL, I. *Outlier* Detection. In: ___. Maimon, O; Rockach, L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. **Kluwer Academic Publishers**, p.1-16, 2005. ISBN 0-387-24435-2. Disponível em: <http://www.eng.tau.ac.il/~bengal/*outlier*.pdf>. Acesso em: 15 maio 2017.

BORGES, R. M. H. **Introdução à Validação de Métodos**. Brasília: CGCRE/DICLA/IMETRO, 2006. 50p. Disponível em: <http://www.inmetro.gov.br/metcientifica/palestras/Renata%20Borges.pdf. Acesso em: 27 maio 017.

CARVALHO, A. T. F. Caracterização climática da quadra chuvosa de município do semiárido brasileiro, entre os anos de 2013 a 2017. **Revista Geografia em Atos**, v. 2, n. 17, p. 4-23, 2020. Disponível em: <https://revista.fct.unesp.br/index.php/geografiaematos/article/view/7116>. Acesso em: 22 nov. 2021.

CROSBY; T. How to detect and handle *outlier*s. **Technometrics**. v.3, n.3, p.315-316. ago., 1994. Disponível em: <ftp://ftp.math.utah.edu/pub/tex/bib/toc/technometrics1990.html#36(3):August:1994>. Acesso em: 15 maio 2017.

DIXON, W. J. Analysis of Extreme Values. Institute of Mathematical Statistics, **The Annals of Mathematical Statistics**, v.21, n.4, p.488-506, 1950. Disponível em: <http://projecteuclid.org/download/pdf_1/euclid.aoms/1177729747>. Acesso em: 27 maio 2017.

FETTERMANN, D. C; GUERRA, K. C; MANO, A. P; MARODIN, G. A. Uma Sistemática para Detecção de Fraudes em Empresas de Abastecimento de Água. **Interciência**, v.40, n.2, p.114-120, 2015.

GRUBBS, F. E. Procedures for detecting outlying observations in samples. **American Statistical Association and American Society for Quality**, v.11, n.11, p.1-21. Fev. 1969. Disponível em: <http://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/*Outlier*Proc_1969.pdf>. Acesso em: 20 maio 2017.

HAWKINS, D. M. **Identification of *Outlier*s**. London: Chapman & Hall, 1980. 188p. Disponível em: <http://professor.ufabc.edu.br/~ronaldo.prati/DataMining/*Outlier*s.%20pdf>. Acesso em: 27 maio 2017.

IGLEWICZ, B; HOAGLIN, D. How to detect and handle *outlier*s. **ASQC Quality Press**, 1993. Disponível em: <http://www.worldcat.org/title/how-to-detect-and-handle-*outlier*s/oclc/901847172?referer=di&ht=edition>. Acesso em: 20 maio 2017.

IDEMA. INSTITUTO DE DESENVOLVIMENTO SUSTENTÁVEL E MEIO AMBIENTE. **Anuário Estatístico do Rio Grande do Norte**. Natal: IDEMA, 2015. Disponível em: <http://www.idema.rn.gov.br/Conteudo.asp?TRAN=ITEM&TARG=1357&ACT=null&PAGE=0&PARM=null&LBL=Socioecon%C3%B4micos>. Acesso em: 12 nov. 2017.

INMET. INSTITUTO NACIONAL DE METEOROLOGIAIDE. **Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP)**. Brasília: INMET, 2015. Disponível em: <http://www.inmet.gov.br/projetos/rede/pesquisa/>. Acesso em: 30 out. 2015.

MAGALHÃES, M. N; LIMA, C. A. P. **Noções de Probabilidade e Estatística**. 7. ed. São Paulo: Editora da Universidade de São Paulo, 2013.

MARENGO, J. A. Vulnerabilidade, impactos e adaptação à mudança do clima no semi-árido do Brasil. **Parcerias Estratégicas**, v. 13, n.27, p.149-176, 2008. Disponível em: <http://seer.cgee.org.br/index.php/parcerias_estrategicas/article/view/329>. Acesso em: 14 ago. 2018.

NIST/SEMATECH. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. Quantitative Techniques - Detection of *Outlier*s. *In:* _____. e-**Handbook of Statistical Methods**: Exploratory Data Analysis (EDA). April, 2012. Disponível em: <http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm>. Acesso em: 27 maio 1981.

PORTAL ACTION. **Boxplot**. Disponível em: <http://www.portalaction.com.br/estatistica-basica/31-boxplot>. Acesso em: 26 abr. 2017a.

PORTAL ACTION. **Dotplot**. Disponível em: <http://www.portalaction.com.br/estatistica-basica/32-dotplot>. Acesso em: 02 maio 2017b.

RAMALHO, M. F. J. L.; GUERRA, A. J. T. O risco climático da seca no semiárido brasileiro. **Territorium**, p.61-74, v. 25, n. 1, 2018. Disponível em: <https://dialnet.unirioja.es/servlet/articulo?codigo=6229238>. Acesso em: 22 nov. 2021.

SEO, S. **A Review and Comparison of Methods for Detecting *Outlier*s in Univariate Data Sets**. 2006. 53p. Dissertação (Mestrado em Saúde Pública) – Programa de Pós-Graduação em Saúde Pública, University of Pittsburgh. 53 p.

SHIFFLER, R. E. Maximum Z Scores and *Outlier*s. **The American Statistician**, v.42, n.1, p.79-80, fev., 1988. Disponível em: <http://wweb.uta.edu/faculty/ricard/Classes/KINE5305/Shiffler%20(1988)%20Maximum%20Z%20scores%20and%20*outlier*s.pdf>. Acesso em: 15 maio 2017.

SILVESTRE, M. R; SANT'ANNA NETO, J. L; FLORES, E. F. Critérios Estatísticos para Definir de Anos Padrão: uma contribuição a Climatologia Geográfica. **Revista Formação**, v.2, n.20, p.23-53, 2013. Disponível em: http://revista.fct.unesp.br/index.php/formacao/article/view/2360/2398. Acesso em: 02 maio 2017.

SILVA, B. C. O. **Precipitação Pluviométrica e Precipitação**: análises do período chuvoso norte-rio-grandense fundamentadas em métodos e técnicas quantitativas. Campina Grande: EPTEC, 2021. 82 p.

SILVA, B. C. O. **As Ablepsias dos Métodos Quantitativos Clássicos**: ênfase na caracterização da quadra chuvosa do Semiárido Potiguar. 2019. 330 p. Tese (Doutorado) - Programa de Pós-Graduação em Geografia, Universidade Federal de Pernambuco (UFPE), Campus de Recife, 2019.

SOUZA, C. L. O.; NOGUEIRA, V. F. B.; NOGUEIRA, V. S. Variabilidade interanual da precipitação em cidades do semiárido brasileiro entre os anos de 1984 e 2015. **Revista Verde de Agroecologia e Desenvolvimento Sustentável**, v. 12, n. 4, p. 740-747, 2017. Disponível em: <https://dialnet.unirioja.es/servlet/articulo?codigo=7161853>. Acesso em: 22 nov. 2021.

TAVARES, V. C.; ARRUDA, I. R. P.; SILVA, D. G. Desertificação, mudanças climáticas e secas no semiárido brasileiro: uma revisão bibliográfica. **Geosul**, v. 34, n. 70, p. 385-405, 2019. Disponível em: <https://periodicos.ufsc.br/index.php/geosul/article/view/2177-5230.2019v34n70p385/38526>. Acesso em: 22 nov. 2021.

TUKEY, J. W. **Exploratory Data Analysis**. 1. ed. Massachusetts: Addison-Wesely, Series in Behavioral Science, 1977. 688p. Disponível em: <http://www.popline.org/node/499313>. Acesso em: 20 maio 2017.

Recebido em setembro de 2021.

Revisão realizada em outubro de 2021.

Aceito para publicação em novembro de 2021