ISSN 2175-6813

Revista Engenharia na Agricultura

Viçosa, MG, DEA/UFV - DOI: 10.13083/reveng.v29i1.11738 V.29, p.157-168, 2021

AN APPROACH TO QUALITY ANALYSIS, GAP FILLING AND HOMOGENEITY OF MONTHLY RAINFALL SERIES

Janaina Cassiano dos Santos[1], Gustavo Bastos Lyra[2], Marcel Carvalho de Abreu[2] & Daniel Carlos de Menezes[3]

1 - Federal Fluminense University, Department of Agricultural and Environmental Engineering, Niterói, Rio de Janeiro, Brazil.

2 - Federal Rural University of Rio de Janeiro, Department of environmental sciences, Seropédica, Rio de Janeiro, Brazil.

3 - Brazilian Airport Infrastructure Company, Rio de Janeiro, Rio de Janeiro, Brazil

Keywords:
Consistency
Linear Regression
Rainfall
Time Series

ABSTRACT

The aim of this work was to propose a method for the consistency of climatic series of monthly rainfall using a supervised and unsupervised approach. The methodology was applied for the series (1961-2010) of rainfall from weather stations located in the State of Rio de Janeiro (RJ) and in the borders with the states of São Paulo, Minas Gerais and Espírito Santo with the State of Rio de Janeiro. The data were submitted to quality analysis (physical and climatic limit and, space-time tendency) and gap filling, based on simple linear regression analysis, associated with the prediction band ($p < 0.05$ or 0.01), in addition to the Z-score (3, 4 or 5). Next, homogeneity analysis was applied to the continuous series, using the method of cumulative residuals. The coefficients of determination (r^2) between the assessed series and the reference series were greater than 0.70 for gap filling both for the supervised and unsupervised approaches. In the analysis of data homogeneity, supervised and unsupervised approaches were effective in selecting homogeneous series, in which five out of the nine final stations were homogeneous ($p > 0.9$). In the other series, the homogeneity break points were identified and the simple linear regression method was applied for their homogenization. The proposed method was effective to consist of the rainfall series and allows the use of these data in climate studies.

Palavras-chave:
Chuva
Consistência
Regressão Linear
Séries Temporais

PROPOSTA DE ABORDAGEM PARA ANÁLISE DE QUALIDADE, PREENCHIMENTO DE FALHAS E HOMOGENEIDADE DE SÉRIES MENSAIS DE PRECIPITAÇÃO

RESUMO

O objetivo foi propor um método para consistência de séries climáticas de precipitação pluvial mensal por meio de abordagem supervisionada e não-supervisionada. O método proposto foi aplicada para séries (1961-2010) de precipitação em estações meteorológicas localizadas no estado do Rio de Janeiro (RJ) e nas divisas com os estados de São Paulo, Minas Gerais e Espírito Santo com o RJ. Os dados foram submetidos à análise de qualidade (limite físico, climático e tendência espaço-temporal) e preenchimento de falhas, baseados na análise de regressão linear simples (RLS), associado à banda de predição ($p < 0,05$ e 0,01), além do escore-Z (3, 4 ou 5). Posteriormente, foi aplicada a análise de homogeneidade às séries contínuas, pelo método do resíduo acumulado. Os coeficientes de determinação (r^2) entre as séries testadas e as séries de referência foram superiores a 0,70 para o preenchimento de falhas, tanto para a abordagem supervisionada, quanto não-supervisionada. Na análise de homogeneidade dos dados, as abordagens supervisionada e não-supervisionada foram efetivas na seleção de séries homogêneas, em que cinco das nove estações finais apresentaram homogeneidade ($p > 0,9$). Nas demais séries foram identificados os pontos de quebra da homogeneidade e aplicado a correção, obtida por RLS para a sua homogeneização. O método proposto foi eficaz para consistir as séries de precipitação e torna possível a utilização desses dados em estudos climáticos.

Received 13 Jan, 2021 • Accepted 11 May, 2021 • Published 16 Aug, 2021

Email (corresponding author)*: jana.cassiano@gmail.com

SECTION EDITOR IN CHARGE
André Pereira Rosa

INTRODUCTION

Precipitation is one of the most important variables in climate studies, with emphasis on climate variability and changes, climatic classification and hydrological modeling, in addition to its importance in the planning, design and management of irrigation and drainage systems and water resources in general (ABREU *et al.,* 2021; JAVARI, 2016). Its occurrence is directly related to water supply, energy generation, agricultural production, human health and wellbeing, natural disasters (droughts, floods, inundation, landslides and mass displacement), among other activities (BRUBACHER *et al.,* 2020).

Climatic studies require long-time, continuous and quality weather observations to better understand the characteristics of the study region. These observations are often from automatic or conventional surface weather or climatological stations (SANTOS et al., 2012). Among the challenges related to rainfall monitoring, the construction and maintenance of a consistent and homogeneous historical database stands out (AY, 2020; SANTOS *et al*., 2012). In Brazil, rainfall series are often restricted, discontinuous, poor quality and not homogeneous in space-time (TOSTES *et al*., 2017).

Data consistency has to be executed so that the results based on these series do not express contradictory and erroneous conclusions. It involves quality analysis, gap filling (when they exist) and testing of the homogeneity of the series (ANA, 2012; ANIMASHAUN *et al*., 2020). The filling is important, due to the difficulties and limitations of several methods to be applied to the non-continuous series, with emphasis on hydrological models, drought indexes and climate variability, trend studies, occurrence probability and return time (DE OLIVEIRA-JÚNIOR *et al*., 2021; LIMA *et al*., 2021; LYRA *et al*., 2006, 2017). In addition, this step can be restricted or have poor accuracy and precision due to physical limitations, such as those caused by complex terrains and/or proximity to the coastal environment (BRUBACHER; *et al.,* 2020). The analysis of data homogeneity after gap filling is rarely performed, incurring in uncertainties about the rainfall data series (CARVALHO; RUIZ, 2016).

In the temporal and spatial scale, the gap filling methods can overlap with each other in terms of accuracy, which reinforces the lack of consensus among authors to list the appropriate methods for each situation. Thus, comparative studies are justified in order to define the best methods and strategies for each variable and conditions (BRUBACHER *et al.,* 2020). Among the methods used to fill gaps in rainfall data, we can mention linear regression (KITE, 1988; PRECINOTO *et al*., 2012), regional weighting and regression weighting (OLIVEIRA *et al*., 2010). However, the regression method stands out for showing better performance in filling rainfall gaps (LO PRESTI *et al.,* 2010; OLIVEIRA *et al*., 2010), being applied in several studies (BERTONI; TUCCI, 2013; CAMERA *et al*., 2014; DE OLIVEIRA-JÚNIOR *et al*., 2012; DI PIAZZA *et al*., 2011; MELLO; SILVA, 2013; NEWMAN *et al*., 2015).

Homogeneity tests allow identifying ruptures in the trend of the series and abrupt changes in the mean and variance of the distribution (SANTOS *et al.,* 2016). Among the methods, double mass, regional vector, cumulative residuals stand out (ALLEN *et al*., 1998; KITE, 1988).

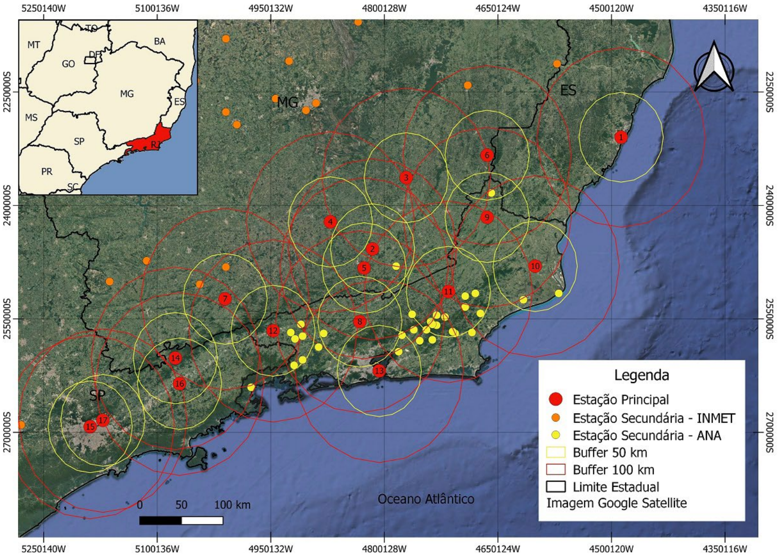
Further, the preliminary analysis of hydroclimatic data is extremely time-consuming and demands attention from the observer at each stage. The advent of software and programming languages leads to the possibility of creating algorithms that facilitate the processing of activities such as quality analysis, gap filling and homogeneity. Therefore, studies that assess supervised and unsupervised rainfall approaches are important. Among these methodologies for gap filling and homogeneity of hydrological series, the linear regression and accumulated residue, respectively, are methods that facilitate an unsupervised approach, in addition to being efficient in their respective objectives (ALMEIDA *et al*., 2021; DE OLIVEIRA-JÚNOR *et al*., 2012; OLIVEIRA *et al*., 2010).

The aim of this work was to i) propose a method for analyzing the quality of rainfall climatic series in a supervised or unsupervised approach and, ii) evaluate the consistency (data quality, gap filling and homogeneity) of historical series of the monthly rainfall, in the period from 1961 to 2010 for the State of Rio de Janeiro (RJ) and locations in the States of São Paulo, Minas Gerais and Espírito Santos, near to the border with the State of Rio de Janeiro, southeastern Brazil.

MATERIAL AND METHODS

Study area and precipitation series

The study region is the state of Rio de Janeiro, southeastern Brazil (latitudes 20°45’54″ and 23°21’57” S and longitudes 40°57’59” and 44°53’18” W, with altitudes between 0 and 2792 m) (Figure 1). The monthly precipitation series were obtained from the Meteorological Database for Teaching and Research - BDMEP, maintained by the National Institute of Meteorology - INMET (Table 1) and by the Hidroweb system, of the National Water Agency - ANA (2012) for the period from 1960 to 2010.



Source: the author

Figure 1. Location map of the study area

Table 1. Data select in the Meteorological Database for Teaching and Research - BDMEP (https://mapas.inmet.gov.br/)

ID	Station	Location	OMM Code	Latitude (graus)	Longitude (graus)	Altitude (m)	Start of the series	End of series	Köppen Climate Classification
1	Vitória	ES	83648	-20.31	-40.31	36.2	Jan/1961	Dec/2010	Am
2	Coronel Pacheco	MG	83037	-21.55	-43.26	435.0	Oct/1966	Abr/2009	Cwb
3	Viçosa	MG	83642	-20.76	-42.86	712.2	Jan/1961	Dec/2010	Cwa
4	Barbacena	MG	83689	-21.45	-43.76	1126.0	Jan/1961	Dec/2010	Cwa
5	Juiz de Fora	MG	83692	-21.76	-43.36	940.0	Jan/1961	Dec/2010	Cwb
6	Caparaó	MG	83639	-20.51	-41.9	843.2	Feb/1973	Dec/2010	Cwb
7	São Lourenço	MG	83736	-22.1	-45.01	953.2	Jan/1961	Dec/2010	Cmb
8	Avelar	RJ	83049	-22.35	-43.41	507.0	Jan/1985	Dec/2010	Cfa
9	Itaperuna	RJ	83695	-21.2	-41.90	123.6	Jan/1961	Dec/2010	Aw
10	Campos dos Goytacazes	RJ	83698	-21.74	-41.33	11.2	Jan/1961	Dec/2010	Aw
11	Cordeiro	RJ	83718	-22.02	-42.36	505.9	Ago/1971	Dec/2010	Cwa
12	Resende	RJ	83738	-22.45	-44.44	439.9	Jan/1961	Dec/2010	Cwa
13	Rio de Janeiro	RJ	83743	-22.89	-43.18	11.1	Jan/1961	Dec/2010	Aw
14	Campos do Jordão	SP	83714	-22.75	-45.60	1642.0	Jan/1961	Dec/2010	Cfb
15	Mirante de Santana	SP	83781	-23.50	-46.61	792.1	Jan/1961	Dec/2010	Cwa
16	Taubaté	SP	83784	-23.03	-45.55	577.0	Jan/1961	Dec/2010	Cfb
17	Guarulhos	SP	83075	-23.43	-46.46	735.0	Jul/1983	Dec/2010	Cfb

The series were initially separated into two groups: one containing the principal weather stations and a second group denominated secondary stations. The principal stations were selected based on: i) location, distance less than 100 km from the border of the state of Rio de Janeiro or located in the state, ii) observation period greater than thirty-five years and iii) with periods of data interruption of less than five consecutive years (10% of the size of the series) (ANA, 2012).

The purpose of the secondary stations was to assist in filling in the months without data in coincident periods and identifying possible errors, for that, they met the criteria (ANA, 2012): i) distance between stations less than 100 km, preferably 50 km, ii) data period longer than twenty years and iii) information on periods absent from the principal stations. Seventeen INMET Weather stations were selected, but only ten met the criteria for principal stations and eighteen from ANA for secondary stations.

The reference series was building using the arithmetic mean of the rainfalls of at least three and at most five other stations (principal and secondary stations) contained around each main station. Firstly, stations with a distance of less than 50 kilometers were chosen, with a maximum limit of 100 kilometers (Figure 1) and coefficient of determination (r^2) greater than 0.50 between the series to have the gaps filled, and each of the stations within the search radius. The averages of rainfall observations from the selected stations were considered representative of the climate trend of the study area.

Quality analysis – supervised and unsupervised method

After the establishment of reference series, for the analysis of data quality of the principal series, two approaches were proposed: a supervised and an unsupervised. The supervised approach was carried out on a step-by-step basis, with criteria to support the observer's decision-making on data quality, while the unsupervised approach was carried out using an algorithm developed in Excel® software that performed the quality analysis and gap filling without the interference of the observer.

The objective of the unsupervised approach is to verify the effectiveness of this algorithm as the preliminary analysis of climate data is extremely time-consuming and slow. However, it is an effective step towards obtaining quality and continuous climate data series.

For the two data quality approaches (supervised and unsupervised), it was fisrtly identified the physical limits of monthly rainfall, that is, limit values that could occur for the evaluated phenomenon, such as monthly rainfall greater than 0 or less than 1000 mm (BRITO *et al.*, 2017). Values beyond this range were considered physically inconsistent and removed from the series. The climatic limits were tested using the Z-score (Equation 1), which checks the relative position of the event, allowing to assess how many standard deviations the event is located far from the mean and the probability of occurrence of an event of such magnitude.

The unsupervised approach sought to identify spurious series values with an interval of three, four and five times the Z-value (Equation 1) of the series, combined with a simple linear regression prediction band with p-value of 0.05 and 0.01. This procedure was performed for each of the principal stations, where the data were considered spurious when simultaneously identified by both methods (greater than the Z limit and beyond the prediction band), which were identified only by one of the steps, were denominated as suspicious.

$$Z = \frac{x - \mu}{sd} \tag{1}$$

Where:

x represents the observed value; μ represents the mean of the observed values and sd represents the standard deviation of the series.

Gap filling

Gap filling was performed with the aid of the simple linear regression (SLR) method. Therefore, the station series and the reference series must meet the criterion of the slope coefficient (β_1) of the simple linear regression ($Y = \beta_0 + \beta_1 X$) to be statistically significant ($\beta_1 \neq 0$) and within the range between $0.7 > \beta 1 > 1.3$ and at the criterion of the coefficient of determination (r^2) greater than 0.70 (ALLEN *et al*., 1998; KITE, 1988; OLIVEIRA *et al*., 2010).

The SLR method consists of estimating data missing in a series, resulting from the linear correlation between the series to be filled and the series of another station, without gaps (reference series). The linear regression ($Y_i = \beta_0 + \beta_1 X_i$) introduces as an estimated variable (Y) the monthly precipitation series of the station under analysis, the predictive variable (X) the reference series of that station (without failures), β_0 represents the intercept (mm) and β_1 is the angular coefficient and the subscript term *i* represents the i-th observation (DOS SANTOS *et al*., 2018).

Homogeneity test

To analyze data homogeneity, the accumulated residue method was applied between the constructed reference series and the value obtained from the arithmetic mean after gap filling using the SLR method. The method of cumulative residuals consists of plotting the residuals of the SLR and, on the same graph, an ellipse defined from the coefficients α, β and θ, where α is the total number of data divided by two, β is obtained through Eq. 2 and θ are the eight equal values in ellipse degrees (0° to 360°).

$$\beta = \frac{(X_i)}{(X_i - 1)\text{^}0.5} * p * Sy \qquad (2)$$

Where:

X_i represents the total number of data, p is the level of significance to be used (90%) and Sy is given by (Eq.3):

$$Sy = \sqrt{\sigma_y * (1 - (r)^2)} \qquad (3)$$

Where: σ_y is the standard deviation of the constructed series, and r is the Pearson correlation coefficient.

When the residual values are beyond the range defined by the ellipse, the series was considered non-homogeneous (KITE, 1988; ALLEN *et al*., 1998). Values outside the prediction range determine the rupture in the data series besides being graphically identified. The series, when not homogeneous, is divided into two subsets: before and after the rupture point.

The homogeneity is corrected using a correction factor (Δ), calculated through the difference between the regression estimates generated from the subsets of the non-homogeneous series (before and after the rupture point). The value Δ is added/subtracted from the original series, considered non-homogeneous, from the rupture point. The procedure is progressively applied to verify the homogeneity of the corrected series, until homogeneity of the data (ALLEN *et al*., 1998).

Analysis of the performance of the supervised and unsupervised methods

The analysis of the performance of the supervised and unsupervised methods was verified based on the ability to identify monthly rainfall series with gaps and inhomogeneities and correct them by means of gap filling based on linear regression and homogeneity through the method of cumulative residuals.

RESULTS AND DISCUSSION

The rainfall time series of all stations used in this work showed some kind of gap in the records (Table 2), where it was selected seventeen stations from INMET, which presented data for the entire analyzed period.

Seven stations were excluded in the initial stage of selection for the data period, due to the low coefficient of determination, distance greater than 100 km from at least three others weather stations and the geopolitical border of the state of Rio de Janeiro as well as the percentage of gaps greater than two-thirds of the series (Table 3).

For these seasons, a large number of gaps were observed particularly in winter (June, July, August) and summer (December, January, February) and from 1984 to 1991.

Regarding the angular coefficient, all ten pre-selected stations had values within the range

Table 2. Percentage of monthly gaps in the assessed stations

Stations	Jan	Fev	Mar	Apr	Mai	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Vitória	18	20	24	28	16	18	18	20	20	18	22	18
Coronel Pacheco	20	20	20	20	24	22	22	24	24	24	18	22
Viçosa	22	18	18	18	18	18	20	20	18	16	18	20
Barbacena	10	14	12	10	10	10	10	8	8	10	8	8
Juiz de Fora	8	8	18	14	10	12	8	6	10	6	6	6
Caparaó	48	40	46	42	46	42	46	42	46	40	44	42
São Lourenço	12	6	8	8	8	8	8	6	6	6	6	6
Avelar	50	52	52	50	50	48	48	52	52	52	50	50
Itaperuna	30	30	28	28	28	26	26	28	28	30	30	30
Campos dos Goytacazes	18	20	20	20	16	16	16	18	18	18	18	20
Cordeiro	48	50	50	50	48	48	46	46	46	46	46	46
Resende	18	20	20	20	18	16	16	20	18	20	20	18
Rio de Janeiro	36	38	38	38	36	36	36	36	36	36	36	36
Campos do Jordão	6	6	6	8	2	8	6	6	10	6	6	6
Mirante de Santana	0	2	2	2	2	2	2	2	2	2	2	0
Taubaté	34	30	30	32	32	34	32	30	30	30	30	30
Guarulhos	46	48	48	48	48	48	46	46	46	46	46	46

Source: Elaborated by the authors

Table 3. Total percentage of gaps, coefficient of determination and series status (Accepted/Rejected)

Stations	Percentage of gaps (%)	r^2	Accepted/Rejected
Vitória	-	-	Rejected (> 100 km)
Coronel Pacheco	21.7	0.76	Accepted
Viçosa	18.7	0.75	Accepted
Barbacena	9.8	0.83	Accepted
Juiz de Fora	9.3	0.76	Accepted
Caparaó	43.66	-	Rejected
São Lourenço	7.3	0.76	Accepted
Avelar	50.5	-	Rejected
Itaperuna	28.5	0.76	Accepted
Campos dos Goytacazes	18.2	0.66	Accepted
Cordeiro	47.5	-	Rejected
Resende	18.7	0.69	Accepted
Rio de Janeiro	36.5	-	Rejected
Campos do Jordão	6.3	0.54	Accepted
Mirante de Santana	-	-	Rejected (> 100 km)
Taubaté	31.2	0.67	Accepted
Guarulhos	46.83	-	Accepted

Source: Elaborated by the authors

determined before and after the analysis. Before the quality analysis, the Taubaté station had the lowest slope ($\beta_1 = 0.71$) and the highest ($\beta_1 = 1.21$) was found in Coronel Pacheco station.

During the quality analysis, spurious values were identified for each method. The difference in the amount of data collected by the supervised and unsupervised methods of data consistency analysis was observed (Table 4).

Regardless of the definition of the methods (supervised/unsupervised), the maximum value of data removed from the gross historical series was 5%. It was possible to observe that the greater the strictness of the parameters (Z = 3 and p < 0.05), with the unsupervised approach, the more data were identified as suspicious and spurious, therefore being removed from the series.

The particularity of the supervised method in relation to the unsupervised method was that some series presented mismatched or suspicious data through the analyses of the prediction bands and Z scores, but could be considered as representative of the local where they were collected by the observer's assessment and thus remained in the series. The unsupervised method was often not effective in identifying some values, such as when the rainfall value indicated zero in months of high rainfall.

Without the interference of the observer, some local peculiarities tend to be ignored, which influences the exclusion of data from the series that could be representative, particularly for studies of extreme events. The loss of this type of information is relevant, especially in countries like Brazil where the number of rainfall stations per unit area tends to be small (ABREU *et al.*, 2021).

For this work, after removing the suspicious observations, the r^2 for the stations increased and remained within the range from 0.63 (Campos do Jordão) to 0.87 (Barbacena), but only nine out of the ten stations met the criterion of $r^2 \geq 0.7$ (Table 5). The r^2 values found in these nine stations express satisfactory data precision, with the exception of the Campos do Jordão station, which presented values below expectations for the selection performed in an unsupervised way.

Table 4. Percentage of data initially identified in each method and removed from the gross series (e) percentage of spurious data and (r) percentage of data removed from the gross series

Stations	Supervised	Z=3 p=0.05		Z=3 p=0.01		Z=4 p=0.05		Z=4 p=0.01		Z=5 p=0.05		Z=5 p=0.01	
		e	r	e	r	e	r	e	r	e	r	e	r
Barbacena	2.3	0.5	4.9	0.3	3.7	*	2.7	*	1.7	*	3.0	*	1.9
Campos dos Goytacazes	4.5	1.2	4.0	0.7	3.1	0.3	2.3	0.3	1.6	0.2	1.1	0.2	1.1
Campos do Jordão	4.9	1.0	2.7	1.0	2.2	0.2	1.9	0.2	2.2	*	2.5	*	2.0
Coronel Pacheco	4.6	0.8	3.0	0.5	2.8	0.3	2.6	0.2	2.0	*	2.6	*	1.3
Itaperuna	2.0	0.7	2.8	0.5	2.7	*	1.0	*	0.5	*	0.8	*	0.7
Juiz de Fora	2.4	0.3	4.9	0.2	4.2	0.2	3.5	0.2	2.2	*	3.2	*	2.4
Resende	3.5	0.3	3.8	0.3	2.1	0.2	3.5	0.2	2.0	*	3.6	*	1.8
São Lourenço	3.7	1.3	5.0	1.0	4.5	0.3	2.4	0.3	2.2	*	3.0	*	1.7
Taubaté	1.0	0.2	2.0	0.2	1.5	*	1.6	*	1.1	*	1.5	*	1.3
Viçosa	2.8	0.7	3.1	0.5	3.3	0.3	2.0	0.3	2.1	*	2.1	*	1.8

Source: Elaborated by the authors

Table 5. Values of the coefficient of determination after the quality step

Stations/ Data removed (%)	Supervised	Z=3 p=0.05	Z=3 p= 0.01	Z=4 p=0.05	Z=4 p=0.01	Z=5 p=0.05	Z=5 p=0.01
Barbacena	0.85	0.85	0.84	0.87	0.86	0.87	0.86
Campos dos Goytacazes	0.73	0.69	0.68	0.71	0.70	0.71	0.72
Campos do Jordão	0.70	0.64	0.63	0.63	0.63	0.65	0.63
Coronel Pacheco	0.85	0.82	0.82	0.82	0.82	0.83	0.82
Itaperuna	0.80	0.76	0.75	0.79	0.78	0.78	0.78
Juiz de Fora	0.82	0.79	0.78	0.81	0.80	0.81	0.80
Resende	0.83	0.83	0.81	0.82	0.80	0.83	0.80
São Lourenço	0.82	0.80	0.80	0.80	0.80	0.81	0.80
Taubaté	0.78	0.75	0.73	0.74	0.72	0.74	0.73
Viçosa	0.78	0.78	0.79	0.81	0.81	0.81	0.80

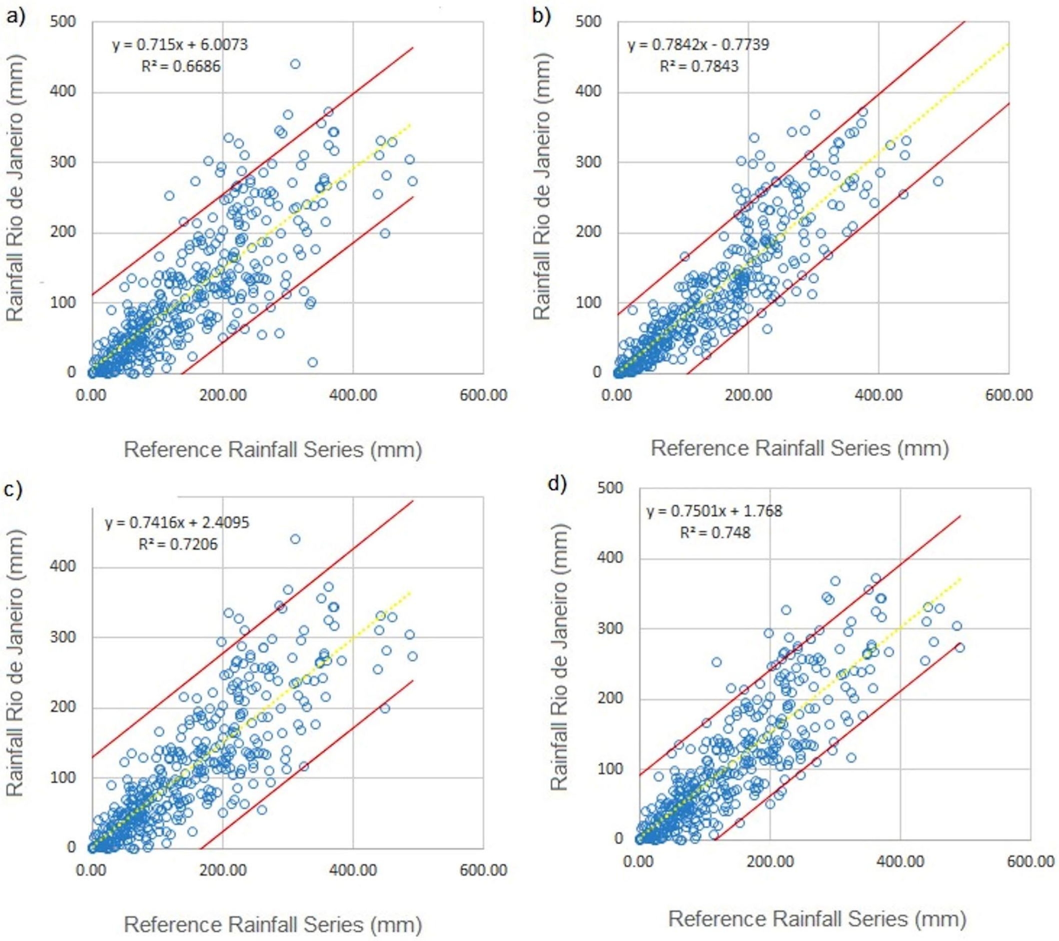
Source: Elaborated by the authors

The stations of Barbacena and Viçosa were the only ones in which the r² values were greater in the unsupervised mode than in manual mode. The station with the greatest discrepancy between the quality assessment methods was the Taubaté station (Figure 2). Values beyond the prediction range were individually evaluated and represent months with rainfall above (below) the monthly climatological average, but because of their climatic importance, they were not removed from the series.

After this process of quality analysis and gap filling, the series were subjected to a homogeneity analysis using the accumulated residue method. Because in 77.78% of the assessed stations the highest r² value identified was given by the supervised method, it was decided to evaluate the homogeneity and trend rupture only in the series that met the criteria of the supervised method.

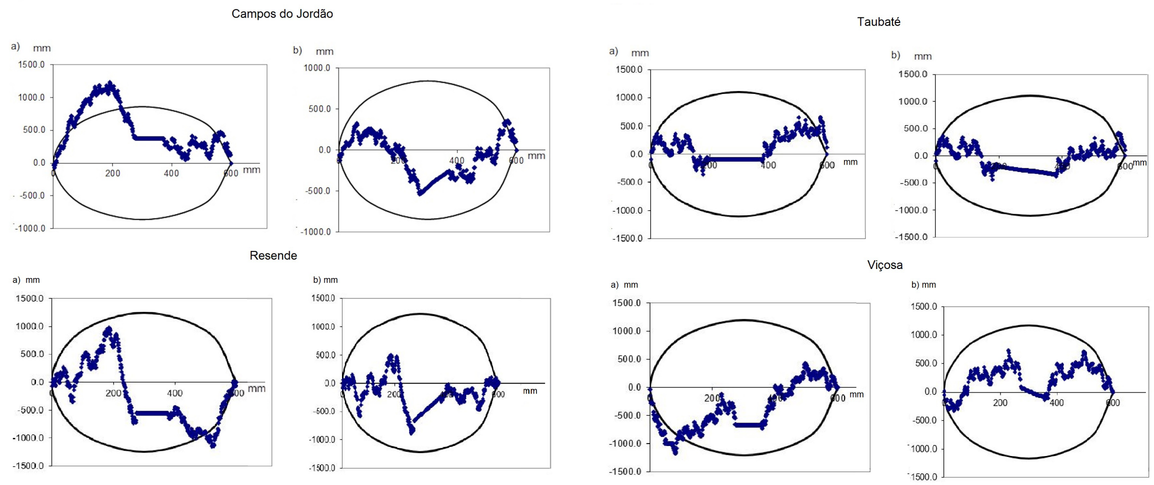
Of the nine selected weather stations, only four (Campos dos Goytacazes, Resende, Taubaté and Viçosa) (Figure 3) showed a rupture in the trend of the series, which represented non-homogeneous data. The other series were homogeneous at a significance level of 90%. It was possible to observe through the graphs the rupture point of each of the four seasons, being December 1975 for Campos dos Goytacazes, January 2008 for Taubaté, November 1969 for Viçosa and January 2003, followed by October 1974 for Resende, the only station in which the procedure was performed twice.

Gois *et al*. (2020) evaluated the normality and homogeneity of a set of rainfall data for the state of Rio de Janeiro, from 1943 to 2013. The authors describe that the highest percentage of gaps was found in stations/series located in the North and Northwest regions of the state of Rio de Janeiro, precisely the regions with lower density of rainfall



Source: Elaborated by the authors

Figure 2. Dispersion of data of the Taubaté station from 1961 to 2010 a) before and b) after manual quality analysis c) after Z = 4 quality analysis and confident level of 99% d) after Z = 3 quality analysis and confidence level of 95%



Source: Elaborated by the authors

Figure 3. Accumulated residue method for data homogeneity analysis for Campos dos Goytacazes station before and after applying the SLR method. a) before homogeneity b) after the homogeneity process

stations (BRITO *et al*., 2017). For the other regions, the data set ranged from 5 to 10% of failures.

In general, the unsupervised method was efficient in the selection of series, whose climatic behavior in the region does not show large variations throughout the year and where there is little availability of nearby weather stations (in the limit of less than 100 km), as there are several factors that may influence this result. The supervised method tends to be more rigorous, as data filtering is done individually and weighted by the observer.

When working with extreme rainfall, the unsupervised method is not recomended because the values that contrast to the Z-score and identified outside the prediction range are automatically removed from the series. Also, these values often represent extreme events in a particular region. Begert *et al*. (2005) identified the increase in the inhomogeneity of rainfall series in the nineteenth century, a change that was justified by the authors to some adjustment in seasonal data, global warming effects and also the introduction of automatic measurement equipment.

The series used in this study, between 1960 and 2010, comprise a period in which trends in the increase in monthly rainfall were detected in the northern region of the state of Rio de Janeiro (SALVIANO *et al.,* 2016), while annual rainfall showed a decrease tendency in regions such as the Baixada Fluminense and Sul Fluminense, between 1951 and 2001 (DERECZYNSKI *et al*., 2013). This fact may have contributed to the number of exclusions of series classified as non-homogeneous, since trends in specific locations cause deviations in the accumulated residual graphs in relation to the average of neighboring locations, without a trend.

In addition, rainfall in the state of Rio de Janeiro has high space-time variability, being influenced by weather systems and orography (ALVARENGA, 2012; CARDOSO; DIAS, 2004; RODRIGUES; WOLLINGS, 2017; SOARES; DIAS, 1986; SOBRAL *et al*., 2019), in addition to anomalies related to climate variability such as El Niño – South Oscillation (ENOS), Atlantic ocean surface temperature, Pacific Decadal Oscillation, South Atlantic Convergence Zone, among others (BARRETO, 2009; DA ROCHA *et al*., 2014; DE OLIVEIRA-JÚNIOR *et al*., 2018; GRIMM, 2003; MINUZZI *et al*., 2007; MOLION, 2003; PRADO *et al*., 2007; SILVA; DERECZYNSKI, 2014; STRECK *et al*., 2009). This spatial variability also influences the behavior of individual series, and may cause deviations in the cumulative residuals graph. Non-homogeneous and trending rainfall series should not be used for frequency analysis or modeling due to the possibility of biased inferences. Therefore, methods that help to select consistent series are extremely important.

CONCLUSIONS

- Based on the criteria established to obtain a satisfactory representation of the temporal series, of the 17 main stations evaluated, only nine stations meet the minimum criteria necessary to carry out all the analyses, which emphasizes the need for creation, maintenance and investment in a database of consistent climate data for the State.

- Through statistical indices, the methods of quality analysis and gap filling of the evaluated rainfall data are adequately precise. The supervised method proves to be more efficient in identifying spurious values and more rigorous in the selection of suspicious data. However, this form of assessment requires knowledge of the observer regarding the removal of data from the series. For that, the unsupervised method, which is more similar to the supervised one, considers as parameters: Z-score = 3 and significance level for the prediction band equal to 95%.
- Most stations are characterized as statistically homogeneous after the application of the methods. Homogeneous series can be used for climate studies and the method is effective in filling the gaps arising from the absence of climate data.

AUTHORSHIP CONTRIBUTION STATEMENT

SANTOS, J.C.: Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft; LYRA, G.B.: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing; ABREU, M.C.: Funding acquisition, Resources, Software, Writing – review & editing; MENEZES, D.C.: Writing – review & editing.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERÊNCIAS BIBLIOGRÁFICAS

ABREU, M.C.; SOUZA, A.; LYRA, G.B.; POBOCIKOVA,I.; CECÍLIO, R.A. Analysis of monthly and annual rainfall variability using linear models in the state of Mato Grosso do Sul, Midwest of Brazil. **International Journal of Climatology**, v. 41, n. S1, 2021.

ALLEN, R.G., PEREIRA, L.S.; RAES, D.; SMITH, M. **Crop evapotranspiration: Guidelines for computing crop water requirements.** Rome: FAO Irrigation and Drainage Paper 56, p. 300, 1998.

ALMEIDA, W.S., SEITZ, S.; OLIVEIRA, L.F.C.; CARVALHO, D.F. Duration and intensity of rainfall events with the same erosivity change sediment yield and runoff rates. **International Soil and Water Conservation Research**, v.9, n.1, p.69-75, 2021.

ALVARENGA, L.A. Precipitação no sudeste brasileiro e sua relação com a Zona de Convergência do Atlântico Sul. **Revista Agrogeoambiental**, v. 4, n. 2, p. 1-7, 2012.

ANA. Orientações para consistência de dados pluviométricos. 2012 Disponível em: <https://arquivos.ana.gov.br/infohidrologicas/cadastro/OrientacoesParaConsistenciaDadosPluviometricos-VersaoJul12.pdf>.

ANIMASHAUN, I.M.; OGUNTUNDE, P.G.; AKINWUMIJU, A.S.; OLUBANJO, O.O. Rainfall Analysis over the Niger Central Hydrological Area, Nigeria: Variability, Trend, and Change point detection. **Scientific Africa**n, v.8, p.e00419, 2020.

AY, M. Trend and homogeneity analysis in temperature and rainfall series in western Black Sea region, Turkey. **Theoretical and Applied Climatology**, v. 139, n. 3-4, p. 837-848, 2020.

BARRETO, N.J. Relação entre oscilação decadal do pacífico, El Niño - Oscilação Sul e a circulação atmosférica de verão na América do Sul. Maceió: Universidade Federal de Alagoas, 2009.

BEGERT, M.; SCHLEGEL, T.; KIRCHHOFER, W. Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000. **International Journal of Climatology**, v. 25, n. 1, p. 65-80, 2005.

BERTONI, J.C.; TUCCI, C.E.M. Ciência e Aplicação. In: UFRGS (Ed.) **Hidrologia**. 4. ed. Porto Alegre: p. 177-241.

BRITO, T.T.; OLIVEIRA-JÚNIOR, J.F.;LYRA, G.B.;GOIS, G.;ZERI, M. Multivariate analysis applied to monthly rainfall over Rio de Janeiro state, Brazil. **Meteorology and Atmospheric Physics**, v. 129, n. 5, p. 469-478, 2017.

BRUBACHER, J.P.; OLIVEIRA, G.G.; GUASSELLI, L.A. Preenchimento de Falhas e Espacialização de Dados Pluviométricos: Desafios e Perspectivas. **Revista Brasileira de Meteorologia**, v. 35, n. 4, p. 615-629, 2020.

CAMERA, C.; BRUGGEMAN, A.; HADJINICOLAOU, P.; PASHIARDIS, S.; LANGE, M.A. Evaluation of interpolation techniques for the creation of gridded daily precipitation (1 × 1 km 2); Cyprus, 1980-2010. **Journal of Geophysical Research: Atmospheres**, v. 119, n. 2, p. 693-712, 2014.

CARDOSO, A.O.; DIAS, P.L.S. Variabilidade Da Tsm Do Atlântico E Pacífico E. **Revista Brasileira de Meteorologia**, v. 19, n. 2, p. 113–122, 2004.

CARVALHO, H.P.; RUIZ, M.V.S. Avaliação da Consistência de Séries Históricas de Chuva da Bacia Hidrográfica do Rio Araguari, em Minas Gerais. **Periódico Eletrônico Fórum Ambiental da Alta Paulista**, v. 12, n. 6, 2016.

DA ROCHA, R.P.; REBOITA, M.S.; DUTRA, L.M.M.; LLOPART, M.P.; COPPOLA, E. Interannual variability associated with ENSO: present and future climate projections of RegCM4 for South America-CORDEX domain. **Climatic Change**, v. 125, n. 1, p. 95-109, 2014.

DE OLIVEIRA, L.F.C.; 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, n. 11, p. 1186-1192, 2010.

DE OLIVEIRA-JÚNIOR, J.F.; LYRA,G.B.; GOIS, G.; BRITO, T.T.; MOURA, N.S.H. Análise de Homogeneidade de Séries Pluviométricas para Determinação do Índice de Seca IPP no Estado de Alagoas. **Floresta e Ambiente**, v. 19, n. 1, p. 101-112, 2012.

DE OLIVEIRA-JÚNIOR, J.F.; GOIS, G.;TERASSI, P.M.B.; JUNIOR, C.A.S.; BLANCO, C.J.C.; SOBRAL, B.S.; GASPARINI,K.A.C. Drought severity based on the SPI index and its relation to the ENSO and PDO climatic variability modes in the regions North and Northwest of the State of Rio de Janeiro - Brazil. **Atmospheric Research**, v. 212, p. 91-105, 2018.

DE OLIVEIRA-JÚNIOR, J.F.; FILHO, W.L.F.C.; SANTIAGO, D.B.; GOIS, G.; COSTA, M.S.; JUNIOR, C.A.S.; TEODORO, P.E.; FREIRE, F.M. Rainfall in Brazilian Northeast via in situ data and CHELSA product: mapping, trends, and socio-environmental implications. **Environmental Monitoring and Assessment**, v. 193, n. 5, p. 263, 2021.

DERECZYNSKI, C.; SILVA, W. L.; MARENGO, J. Detection and Projections of Climate Change in Rio de Janeiro, Brazil. **American Journal of Climate Change**, v. 02, n. 01, p. 25-33, 2013.

DI PIAZZA, A.; CONTI, F. L.O.; NOTO, L.V.; VIOLA, F; LA LOGGIA, G. Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. **International Journal of Applied Earth Observation and Geoinformation**, v. 13, n. 3, p. 396-408, 2011.

GRIMM, A. M. The El Niño Impact on the Summer Monsoon in Brazil: Regional Processes versus Remote Influences. **Journal of Climate**, v. 16, n. 2, p. 263-280, 2003.

GOIS, G; DE OLIVEIRA-JÚNIOR, J.F.; DA SILVA JUNIOR, C. A.; SOBRAL, B.S.; DE BODAS TERASSI, P.M.; JUNIOR, A.H.S.L. Statistical normality and homogeneity of a 71-year rainfall dataset for the state of Rio de Janeiro—Brazil. **Theoretical and Applied Climatology**, v. 141, n. 3-4, p. 1573-1591, 2020.

JAVARI, M. Trend and Homogeneity Analysis of Precipitation in Iran. **Climate**, v. 4, n. 3, p. 44, 2016.

KITE, G.W. Frequency and risk analyses in hydrology. Littleton: Water Resourcers Publications, 1988.

LIMA, A.O.; LYRA,G.B.; ABREU, M.C.; DE OLIVEIRA-JÚNIOR, J.F.; CUNHA-SERI,G.; ZERI, M. Extreme rainfall events over Rio de Janeiro State, Brazil: Characterization using probability distribution functions and clustering analysis. **Atmospheric Research**, v.247, p.105221, 2021.

LO PRESTI, R.; BARCA, E.; PASSARELLA, G. A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). **Environmental Monitoring and Assessment**, v. 160, n. 1-4, p. 1-22, 2010.

LYRA, G.B.; GARCIA, B.I.L.; PIEDADE, S.M.S.; SEDIYAMA, G.C.; SENTELHAS, P.C. Regiões homogêneas e funções de distribuição de probabilidade da precipitação pluvial no Estado de Táchira, Venezuela. **Pesquisa Agropecuária Brasileira**, v. 41, n. 2, p. 205-215, fev. 2006.

LYRA, G.B.; DE OLIVEIRA-JÚNIOR, J.F.; GOIS, G.; CUNHA-SERI,G.; ZERI, M. Rainfall variability over Alagoas under the influences of SST anomalies. **Meteorology and Atmospheric Physics**, v. 129, n. 2, p. 157-171, 2017.

MELLO,C.R.; SILVA, A.M. Hidrologia:Princípios e aplicações em sistemas agrícolas. Lavras: UFLA, 2013.

MINUZZI, R.B., SEDIYAMA, G.C.; COSTA, J.M.N.; VIANELLO, R.L.; Influência da La Niña na estação chuvosa da região sudeste do Brasil. **Revista Brasileira de Meteorologia**, v. 22, n. 3, p. 345-353, 2007.

MOLION, L.C.B. Aquecimento global , El Niño , Manchas Solares , Vulcões e Oscilação Decadal do Pacífico. **Climanálise**, n. 1, p. 2-6, 2003.

NEWMAN, A.J.; CLARK, M.P.; CRAIG, J.;NIJSSEN, B.; WOOD, A.; GUTMANN, E.; MIZUKAMI, N.; BREKKE, L.;ARNOLD, J.R. Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States. **Journal of Hydrometeorology**, v. 16, n. 6, p. 2481-2500, 2015.

PRADO, L.F. FILHO, A.J.; LOBO, G.A.; HALLAK, R. Variabilidade Espaço-Temporal Da Precipitação No Estado De São Paulo E Sua Relação Com Enos Entre 1947 E 1997. **XVII Simpósio Brasileiro de Recursos HÌdricos**, p. 1-15, 2007.

RODRIGUES, R.R.; WOLLINGS, T. Impact of Atmospheric Blocking on South America in Austral Summer. **Journal of Climate**, v. 30, n. 5, p. 1821-1837, 2017.

PRECINOTO, R.S.; CORREIA, T.P.; SANTOS, E. O.; LYRA, G.B. Aplicação de regressão linear múltipla para preenchimento de falhas de dados pluviométricos no estado do Rio de Janeiro. XVII Congresso Brasileiro de Meteorologia. Gramado: Sociedade Brasileira de Meteorologia - SBMET, 2012.

SALVIANO, M.F.; GROPPO, J.D.; PELLEGRINO, G.Q. Análise de Tendências em Dados de Precipitação e Temperatura no Brasil. **Revista Brasileira de Meteorologia**, v. 31, n. 1, p. 64-73, 2016.

SANTOS, R.S.; SEDIYAMA, G.C.; OLIVEIRA, R.A.; ABRAHÃO, G.M. Homogeneidade de séries climatológicas em Minas Gerais. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 16, n. 12, p. 1338-1345, 2012.

SANTOS, S.R.Q.; SANSIGOLO, C.A.; SANTOS, A.P.P. Dinâmica temporal em múltiplas escalas de tempo dos eventos secos e chuvosos no sudeste do Brasil. **Revista Brasileira de Geografia Física**, v. 9, n. 5, 2016.

SILVA, W.L.; DERECZYNSKI, C.P. Caracterização Climatológica e Tendências observadas em extremos climáticos no Estado do Rio de Janeiro. **Anuario do Instituto de Geociencias**, v. 37, n. 2, p. 123-138, 2014.

SOARES, J.R.; DIAS, M.A.F.S. Probabilidade de ocorrência de alguns eventos meteorológicos extremos na cidade de São Paulo. **Revista Brasileira de Meteorologia**, v. 1, p. 67-75, 1986.

SOBRAL, B.S.; DE OLIVEIRA-JÚNIOR, J.F.; GOIS,G.; PEREIRA-JÚNIOR, E.R.; TERASSI, P.M.B.; LYRA, G.B.; ZERI, M. Drought characterization for the state of Rio de Janeiro based on the annual SPI index: trends, statistical tests and its relation with ENSO. **Atmospheric Research**, v. 220, p. 141-154, 2019.

STRECK, N.A., BURIOL, G.A.; HELDWEIM, A.B.; GABRIEL, L.F.; DE PAULA, G.M. Associação da variabilidade da precipitação pluvial em Santa Maria com a Oscilação Decadal do Pacífico. **Pesquisa Agropecuaria Brasileira**, v. 44, n. 12, p. 1553-1561, 2009.

TOSTES, J.O.; LYRA, G.B.; OLIVEIRA- JÚNIOR, J.F.; FRANCELINO, M.R. Assessment of gridded precipitation and air temperature products for the State of Acre, southwestern Amazonia, Brazil. **Environmental Earth Sciences**, v. 76, n. 4, p. 153, 2017.