ISSN: 2447-3359

REVISTA DE GEOCIÊNCIAS DO NORDESTE

Northeast Geosciences Journal

v. 11, nº 1 (2025)

https://doi.org/10.21680/2447-3359.2025v11n1ID33963

Evaluation of the influence of natural variables on air temperature through linear regression

Avaliação da influência de variáveis naturais na temperatura do ar através de regressão linear

Letícia Cardoso Anastacio[1]; William Rodrigo Dal Poz[2]; Nilcilene das Graças Medeiros[3] Afonso de Paula dos Santos[4]

[1] Federal University of Viçosa, Civil Engineering Department, Viçosa/MG, Brasil. Adress:leticia.anastacio@ufv.br
ORCID: https://orcid.org/0000-0003-0245-6880

[2] Federal University of Viçosa, Civil Engineering Department, Viçosa/MG, Brasil. Adress: william.dalpoz@ufv.br
ORCID: https://orcid.org/0000-0001-9532-3643

[3] Federal University of Viçosa, Civil Engineering Department, Viçosa/MG, Brasil. Adress: nilcilene.medeiros@ufv.br
ORCID: https://orcid.org/0000-0003-0839-3729

[4] Federal University of Viçosa, Civil Engineering Department, Viçosa/MG, Brasil. Adress: afonso.santos@ufv.br
ORCID: https://orcid.org/0000-0001-7248-4524

Abstract: Air temperature is a factor that directly affects the physiological processes of plants and animals. Therefore, its understanding has become essential in analyzing the adaptations of fauna and flora in certain areas, as well as in studies related to human thermal comfort. However, a major obstacle to this territorial characterization is the lack of a dense network for monitoring climatic data. Various methods have been applied to estimate temperatures in areas without meteorological data, including multiple linear regression equations. This study aims to create a model to predict the average air temperature based on altitude, latitude, longitude, and surface temperature obtained from orbital sensors. The choice of the study area encompassed the states of the South, Southeast, and Mato Grosso do Sul, with 319 weather stations. The resulting regression model achieved an R^2 of 0.74. Temperature predictions and residuals were generated, highlighting higher temperatures in the Midwest, demonstrating the influence of the continentality phenomenon, and lower temperatures in mountainous areas due to altitude. This study underscores the importance of climate monitoring in understanding geographical influences on air temperatures.

Keywords: Temperature; Linear regression; Orbital Image.

Resumo: A temperatura do ar é um fator que tem efeito direto nos processos fisiológicos de plantas e animais. Assim, seu conhecimento tornou-se fundamental na análise das adaptações da fauna e flora em determinadas áreas, assim como em estudos relacionados ao conforto térmico do ser humano. No entanto, um grande obstáculo para essa caraterização territorial é a falta de uma rede densa de monitoramento de dados climáticos. Diferentes métodos têm sido aplicados para estimar temperaturas em áreas sem dados meteorológicos, incluindo equações de regressão linear múltipla. Este estudo visa criar um modelo para prever a temperatura média do ar com base em altitude, latitude, longitude e temperatura de superfície obtida de sensores orbitais. A escolha da área de estudo abrangeu os estados do Sul, Sudeste e Mato Grosso do Sul, com 319 estações meteorológicas. O modelo de regressão resultante alcançou um R^2 de 0,74. Predições de temperatura e resíduos foram gerados, destacando altas temperaturas no Centro-Oeste demonstrando a influência do fenômeno de continentalidade e temperaturas mais baixas em áreas montanhosas devido à altitude. Este estudo destaca a importância do monitoramento climático para entender as influências geográficas nas temperaturas do ar.
Palavras-chave: Temperatura; Regressão linear; Imagem orbital.

Received: 20/09/2023; Accepted: 10/04/2024; Published: 24/02/2025.

1. Introduction

The air temperature is one of the climatic factors that has the greatest direct and significant effect on many physiological processes in plants and animals, including humans. Thus, its knowledge has become fundamental in analyzing the adaptations of plants and animals to specific areas (MEDEIROS et al., 2005).

According to Capuchinho et al. (2019), different methods have been applied to achieve greater accuracy in air temperature estimates where there is no meteorological data, in order to overcome the limitations resulting from the lack of these climatic records. Methods such as the application of neural networks and the creation of regression models can be mentioned. Regression analysis is used to estimate this information because it is a practical, efficient methodology that does not require large computational capacity.

The adjustment of multiple regression equations allows estimating the mean values of temperature, based on longitude, latitude, altitude, and other environmental variables. The estimation of minimum, mean, and maximum monthly and annual air temperatures by geographical coordinates has been the subject of many studies in different states and regions of Brazil. For example, a study by Capuchinho et al. (2019) estimated maximum, minimum, and mean air temperatures for the state of Goiás using geographical coordinates, altitude, and air temperature data obtained between 1987 and 2017. The results indicated that the means of maximum and mean temperatures can be satisfactorily estimated based on altitude, latitude, and longitude, with adjusted coefficients of determination (R^2) ranging from 0.66 to 0.74 for mean temperature, 0.66 to 0.82 for maximum temperature, and 0.56 to 0.72 for minimum temperature. The estimation of meteorological data is crucial when observational data is not available, but the reliability of these estimates depends on the agreement between predicted and observed data.

Lima et al. (2020) estimated temperature by correlating satellite images with monthly mean air temperature data between 1981 and 2010 in the state of Ceará, obtaining a correlation coefficient of 0.96, demonstrating the method's efficiency. Paula et al. (2021) estimated air temperature in the municipality of Cuiabá, MT, using neural networks, with air temperature, humidity, and land cover as variables. The results showed that the models presented a high relationship between measured and predicted data, with R^2 ranging between 0.88 and 0.99. The differences between measured and estimated air temperature data were small, indicating good model accuracy.

In this context, the present study aims to develop a multiple linear regression model to estimate air temperature values throughout the study area using latitude, longitude, surface temperature, and altitude variables, as well as to evaluate the relationship of these variables with air temperature.

The study area of this work includes the states of the South, Southeast, and Mato Grosso do Sul, covering an area of 1,858,410 km² and approximately 319 meteorological stations monitored by INMET (National Institute of Meteorology). It is relevant to highlight that the data used in the model training refer to August 8, 2021. The choice to use data from only one day to estimate air temperature is due to the difficulty in obtaining high-quality orbital images for the entire study region without cloud cover. This limitation is common in extensive areas such as the states covered in this work. According to Filgueiras et al. (2016), the large distance between meteorological network stations hinders the study of thermal conditions in regions where temperature data collection is not performed. The correlation between data collected by stations and data estimated through remote sensing promotes the possibility of determining estimated air temperature values in areas where data availability is lacking.

2. Methodology

2.1 Materials

The following materials were used in the present study:

- Orbital image from the MODIS sensor, aboard the AQUA satellite, with a spatial resolution of 1km;
- Hourly temperature data from INMET meteorological stations in the region of interest;
- R and QGIS software.

2.2 Study Area

The study area encompasses states in the South, Southeast, and Midwest regions of Brazil (Figure 1).

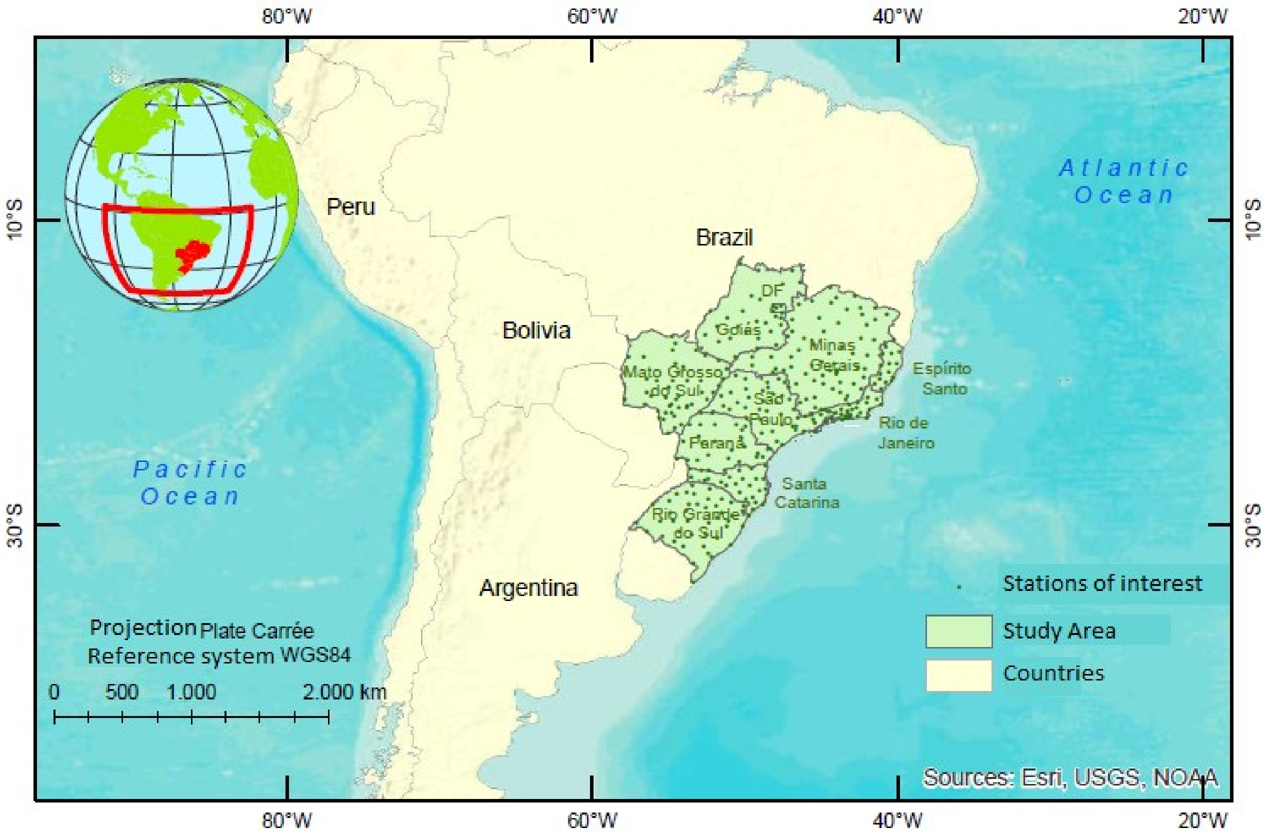

Figure 1 – Study Area.
Source: Authors (2023).

The states represented in Figure 1 have varied climates. Those composing the Southern region have predominantly temperate climates, characterized by well-defined seasons, evenly distributed rainfall throughout the year, hot summers, and pronounced winters. The other states have a Central Brazilian Tropical climate, with rainy and hot summers, and dry winters with moderate cold (NIMER, 1989), as illustrated in Figure 2.

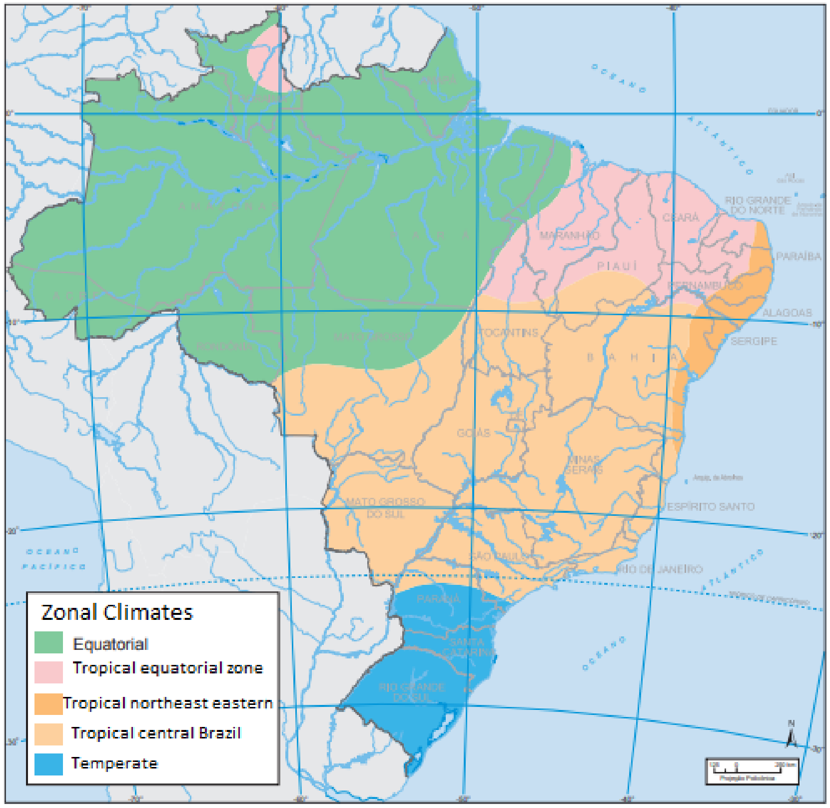

Figure 2 – Climatic Classification of Brazil.
Source: Nimer (1979).

2.3 Surface Temperature Estimation

For the estimation of surface temperature, an image from the AQUA satellite, MODIS sensor, dated August 8, 2021, was used. The scale factor was applied to convert the image to Kelvin temperature (WAN, 2019) (Equation 1).

$$T_k = DN * 0{,}02 \tag{1}$$

Where:
T_k is the temperature in Kelvin scale;
DN is the digital number of the pixel.
Then, the Kelvin scale was converted to degrees Celsius (Equation 2).

$$T_{°C} = T_k - 273{,}15 \tag{2}$$

Where:
$T_{°C}$ is the temperature in degrees Celsius.

2.4 Average air temperature and altitude

The air temperature and altitude data are from INMET meteorological stations distributed in the region of interest. 319 stations were used as illustrated in Figure 1. The daily average for the air temperature variable for August 8, 2021, was calculated for all stations. This day was selected due to low cloud cover over the study region. A boxplot graph was generated, and outlier analysis was performed. In this analysis, it was detected that some stations had readings for only a few hours of the day, so 6 stations were excluded from the analysis. Subsequently, a Shapiro-Wilk normality test was conducted. Finally, correlation indices were calculated to assess how the variables used relate to the average temperature.

2.5 Multiple linear regression

Multiple linear regression is a multivariate technique aimed at obtaining a mathematical relationship between one of the variables under study (dependent variable) and the remaining variables that describe the system (independent variables). Its main application, after discovering the mathematical relationship, is to generate values for the dependent variable when the independent variables are known (LAPPONI, 2005).

For the application of a linear regression model, the variables need to meet the linearity assumption, i.e., the relationship between the dependent and independent variables should be linear (MONTGOMERY et al., 2012). Therefore, an analysis was conducted to evaluate the need to apply transformations to the dependent variable to make its relationship with the other variables linear. In the present case, it was necessary to apply the square root to the average temperature variable. The resulting model is represented in Equation 3.

$$\sqrt{T_{média}} = a_1 * alt + a_2 * lat + a_3 * lon + a_4 * T_s + C \quad (3)$$

Where:
$T_{média}$ is the average air temperature for the station in degrees Celsius;
Alt is the altitude for each station in meters;
Lat is the latitude for each station;
Lon is the longitude for each station;
T_s is the surface temperature in degrees Celsius;
C is the intercept;
$a_{1,}\, a_{2,}\, a_{3,}\, a_4$ are the coefficients to be obtained.

After regression analysis, student t-tests were conducted to assess whether the variables had coefficients statistically different from zero. An analysis of residuals was also performed, evaluating autocorrelation, skewness, kurtosis, heteroscedasticity, and normality in the data.

3. Results and discussion

3.1 Exploratory analysis

In the exploratory data analysis, Pearson correlation coefficients were calculated to assess the relationship between altitude and surface temperature variables, outliers were identified, and a hypothesis test was conducted to evaluate the normality of the data.

Figures 3, 4, 5, and 6 illustrate the correlations between the independent variables and the dependent variable, as well as density plots for each variable.

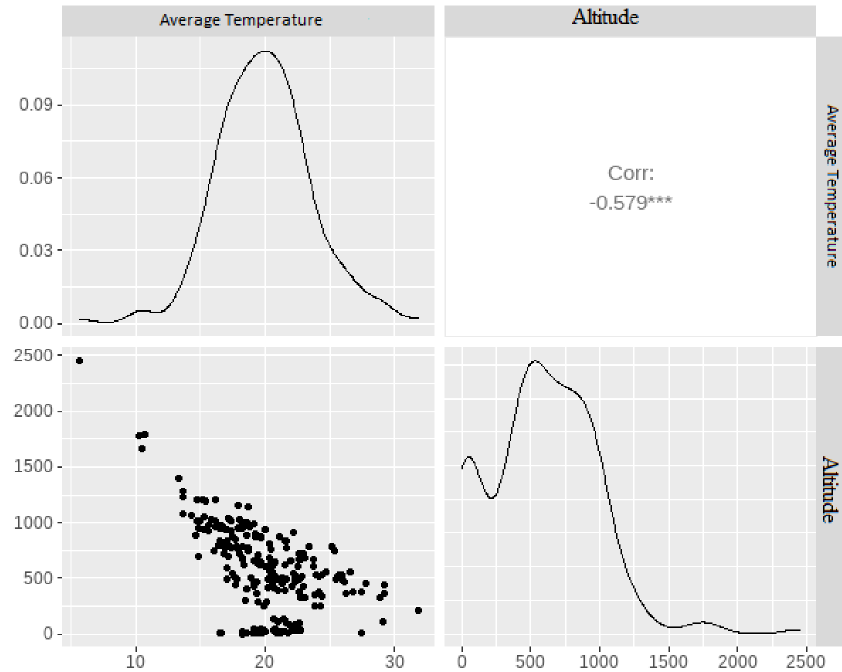

Figure 3 – Density and correlation plots between average temperature and altitude. Source: Authors (2023).

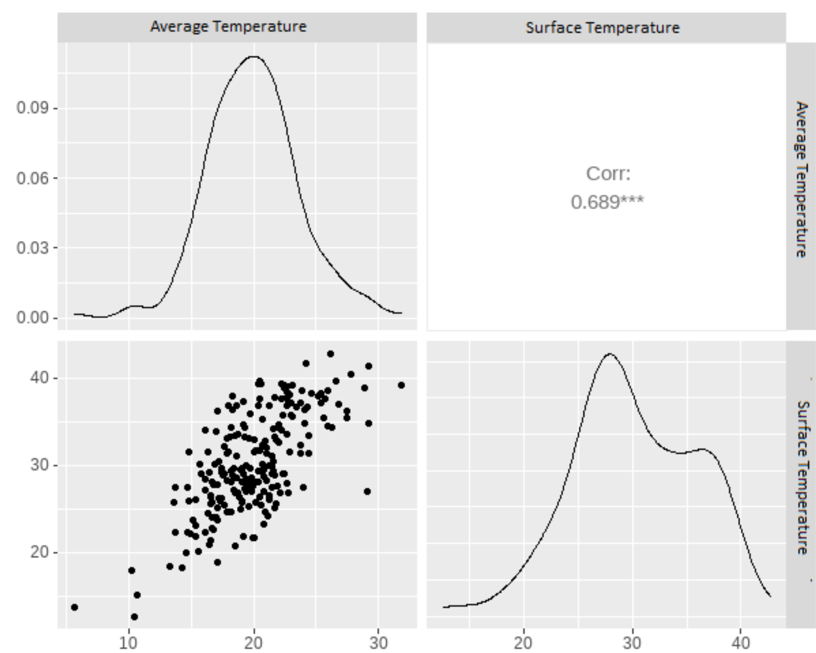

Figure 4 – Density and correlation plots between average temperature and surface temperature. Source: Authors (2023).

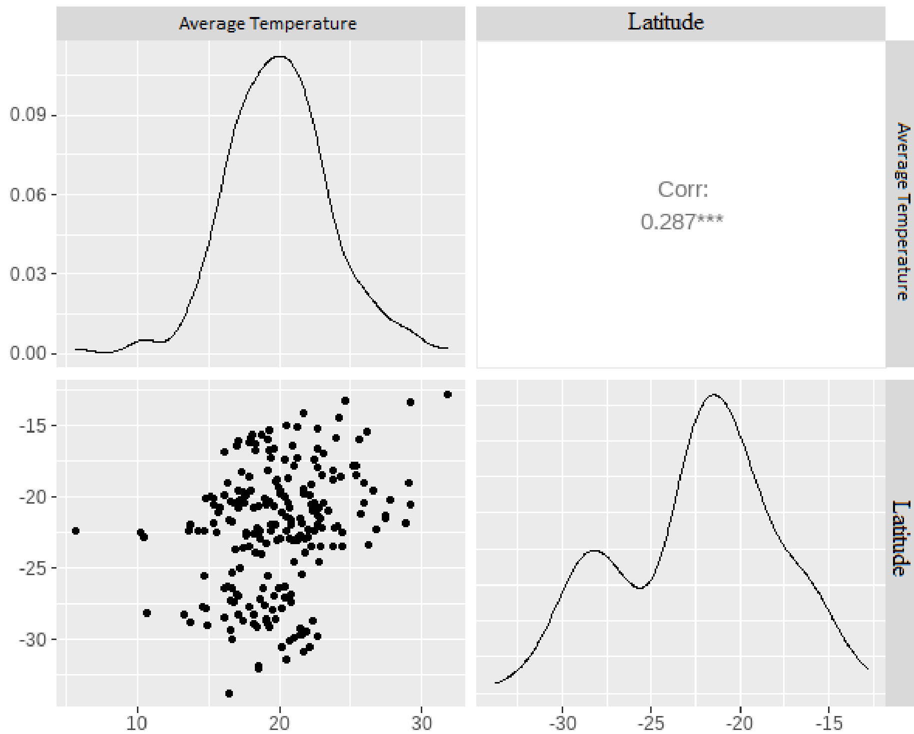

Figure 5 – Density and correlation plots between average temperature and latitude. Source: Authors (2023).

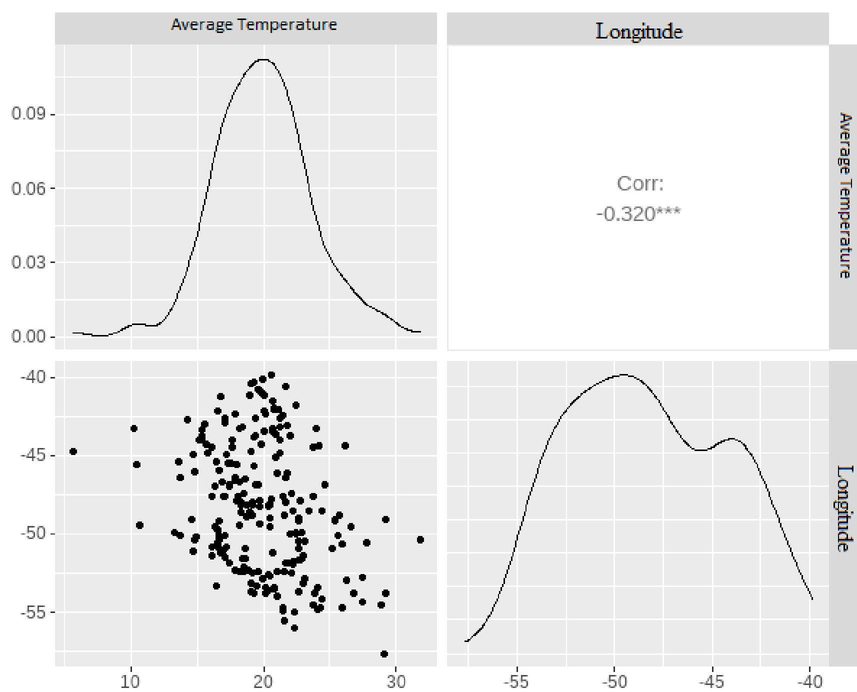

Figure 6 – Density and correlation plots between average temperature and longitude. Source: Authors (2023).

It is observed that altitude and longitude variables have strong and moderate negative correlations, respectively. Surface temperature showed a strong positive correlation, and latitude showed a moderate positive correlation with average air temperature.

A Shapiro-Wilk test was also conducted to assess the normality of the average temperature variable, yielding a p-value of 0.45 with a significance level of 5%, indicating that the data are normally distributed.

3.2 Surface Temperature

In choosing the image, the minimum cloud cover possible was considered to avoid significant interference with the results. Figure 7 illustrates the surface temperatures obtained from the orbital image in the states of interest.

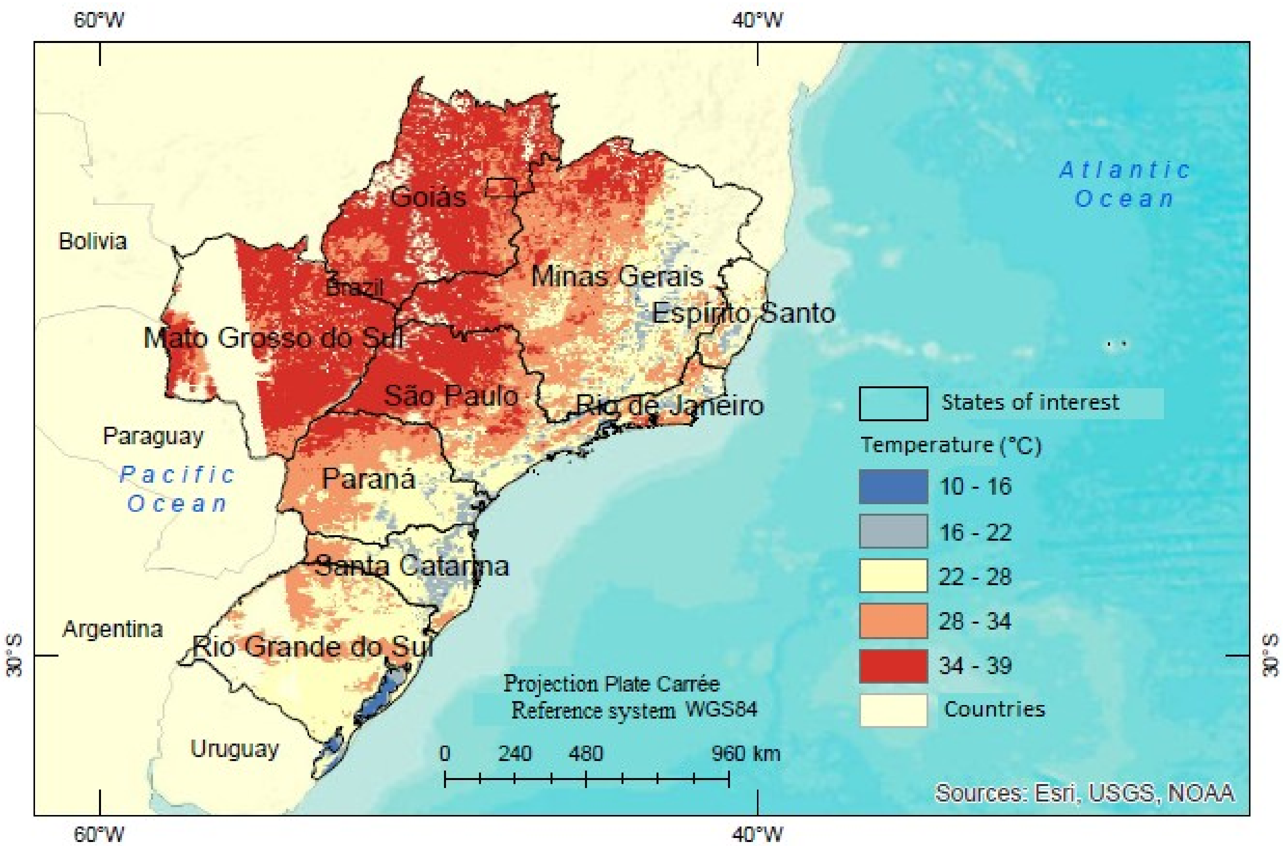

Figure 7 – Surface temperatures obtained from orbital image.
Source: Authors (2023).

It can be observed that the highest temperatures are found in the Midwest region, Western São Paulo, and Minas Gerais. The milder temperatures are concentrated closer to the coast. The white regions represent areas where it was not possible to obtain temperature values due to cloud cover. Stations located in such areas were excluded from the regression analysis.

3.3 Multiple linear regression

The model obtained in multiple linear regression is described in Equation 4.

$$\sqrt{T_{média}} = -0{,}0006602 * alt + 0{,}03824 * lat + -0{,}04193 * long + 0{,}01974 * T_s + 3{,}07 \quad (4)$$

The regression coefficients are exposed in Table 1, along with the p-values obtained for each variable.

Table 1 – Coefficients and P values obtained in multiple linear regression.

Variables	**Coefficients**	**P values**
Altitude	-0,0006602	2E-16
Latitude	0,03824	1,42E-11
Longitude	-0,04193	5,58E-16
Surface Temperature	0,04193	5,39E-7
Intercept	3,07	2E-16

Source: Authors (2023).

The R² coefficient obtained in the model was 0.79, indicating that the model explains approximately 79% of the response variable. All variables obtained a p-value below 0.05 in the student t-test (Table 1), meaning they are considered significant in explaining the dependent variable.

From the positive coefficients of latitude and negative coefficients of longitude, it was possible to identify a trend of increasing air temperature from south to north towards the equator and from east to west, i.e., from the ocean to the continent. The elevation of temperature towards the equator can be explained by the increase in solar radiation balance that occurs from the poles to the equator, due to the inclination of incident solar rays, which is more pronounced in regions near the equator. The increase in temperature towards the continent (east-west direction) results from the attenuation of solar radiation and the phenomenon of continentality, being more evident in coastal areas and less noticeable in arid and semi-arid areas. Regions near oceans and large lakes tend to have higher cloud cover and air humidity, while coastal regions have lower levels of these variables. The higher the cloud cover and air humidity, the greater the atmospheric attenuation, resulting in a lower amount of radiant energy reaching the surface and, consequently, lower air temperatures (MENDONÇA and OLIVEIRA, 2007).

Table 2 illustrates the influence of each variable on air temperature. It can be observed that altitude explained 37% of the variation in air temperature, while surface temperature explained 48%. These were the most influential variables observed in multiple linear regression analysis, also being variables that showed a strong correlation with air temperature.

Table 2 – Contribution of each variable to air temperature.

Variables	**R²**	**Influence (%)**
Altitude	0,3719	37,19%
Latitude	0,07098	7,098%
Longitude	0,09154	9,154%
Surface Temperature	0,4801	48,01%

Source: Authors (2023).

The effect of proximity to the Equator, where solar rays strike more directly on the surface, is represented by the latitude variable, showing the least influence on air temperature. The effect of continentality, in this case represented mainly by longitude, showed the second least influence, indicated by the increase in R².

Figure 8 illustrates a plot of observed values (x-axis) contrasted with values predicted by the model (y-axis). The closer the blue line to the red line, the more explanatory the model can be considered.

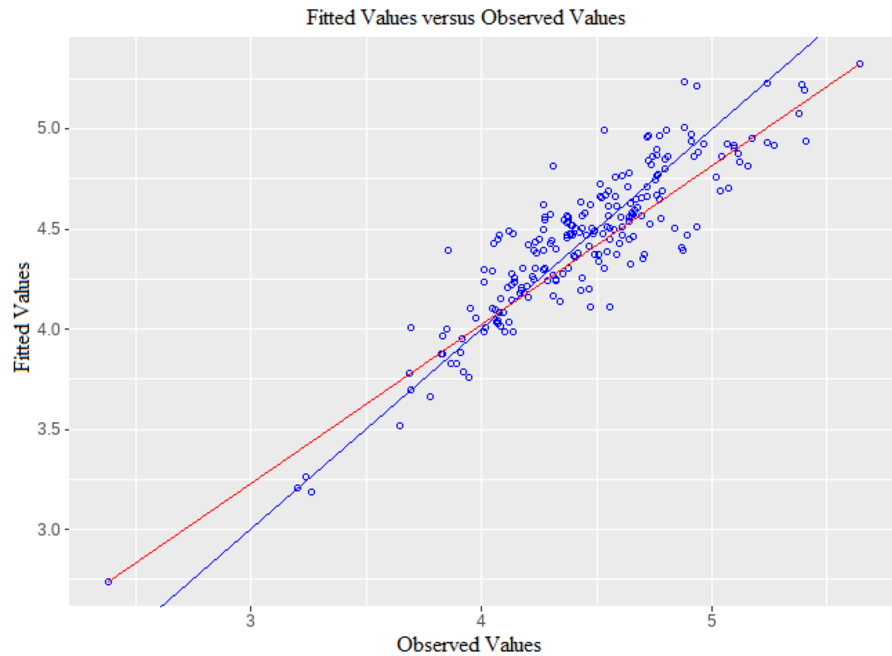

Figure 8 – Plot of observed values (x-axis) versus adjusted values (y-axis).
Source: Authors (2023).

It can be observed that the two lines are close, suggesting a good fit of the regression equation obtained.

3.4 Residuals analysis

Figure 9 illustrates the histogram of residuals, and Figure 10 shows a Q-Q plot, both were used for an initial assessment of the normality of residuals.

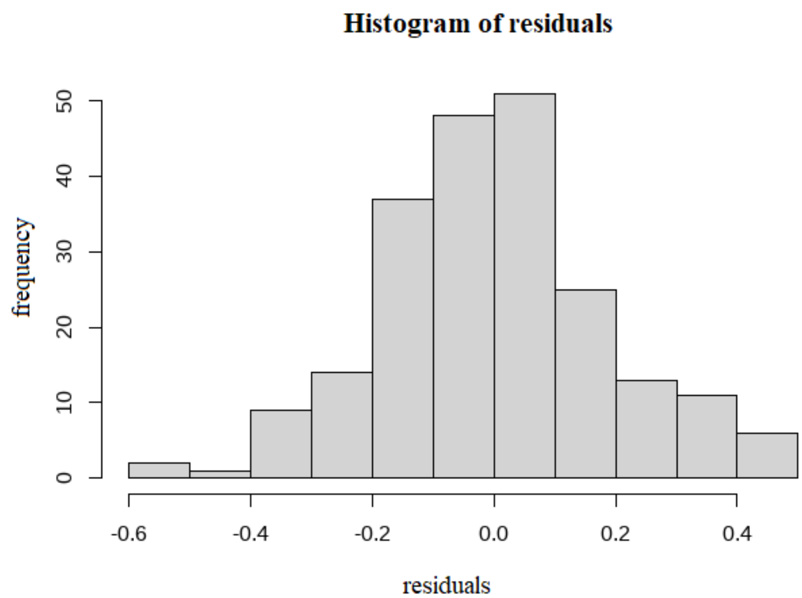

Figure 9 – Histogram of residuals.
Source: Authors (2023).

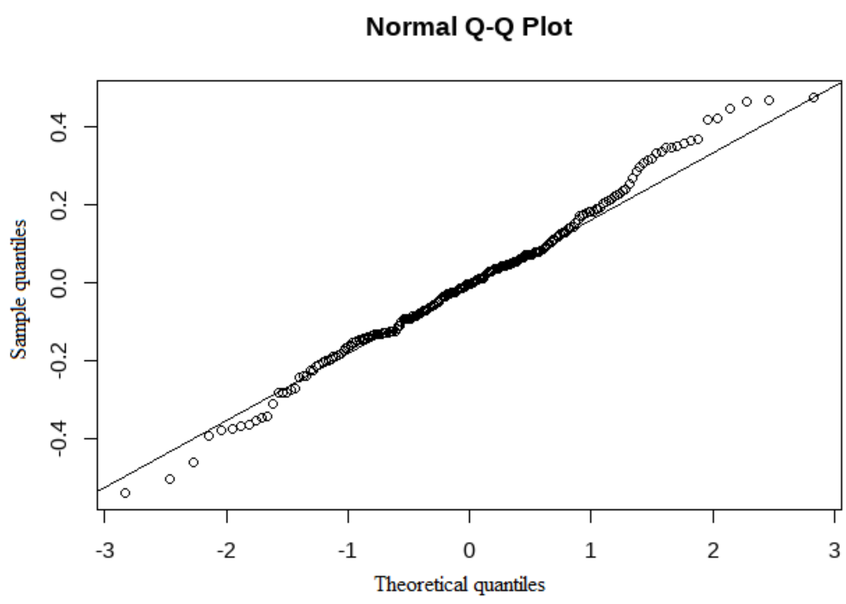

Figure 10 – Q-Q Plot of residuals.
Source: Authors (2023).

The graphs in Figures 9 and 10 indicate that the residuals have a normal distribution. However, for a more detailed analysis, a Shapiro-Wilk test was performed to analyze the normality of residuals, yielding a p-value of 0.05, therefore it was considered that the residuals follow a normal distribution.

For autocorrelation analysis, a Durbin-Watson test was applied, yielding a p-value of 0.43, therefore, no autocorrelation was detected in the residuals.

3.5 Estimated air temperature

Finally, a figure representing the temperatures estimated by the fitted model was generated (Figure 11).

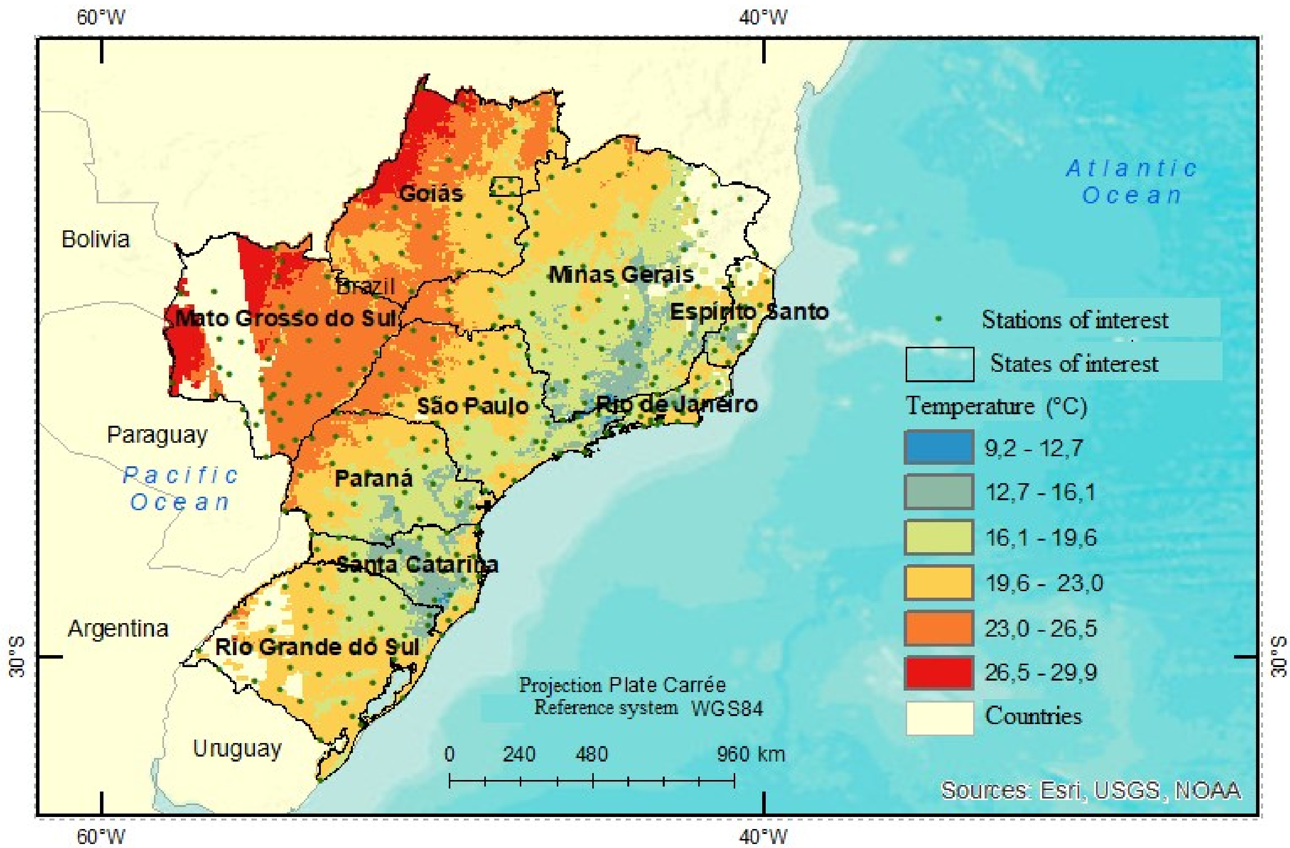

Figure 11 – Estimated temperatures.
Source: Authors (2023).

Regions with lower temperatures are regions of high altitude, such as the Mantiqueira mountain range, located on the border between Minas Gerais, São Paulo, and Espírito Santo. The highest temperatures are concentrated in the Midwest region, which can be explained by the action of continentality, which are climatic effects that affect areas far from the coast, often causing high temperature values and low air humidity (MENDONÇA and OLIVEIRA, 2007).

The difference between the predicted and observed values was also calculated, and an interpolated residual figure was generated through kriging (Figure 12).

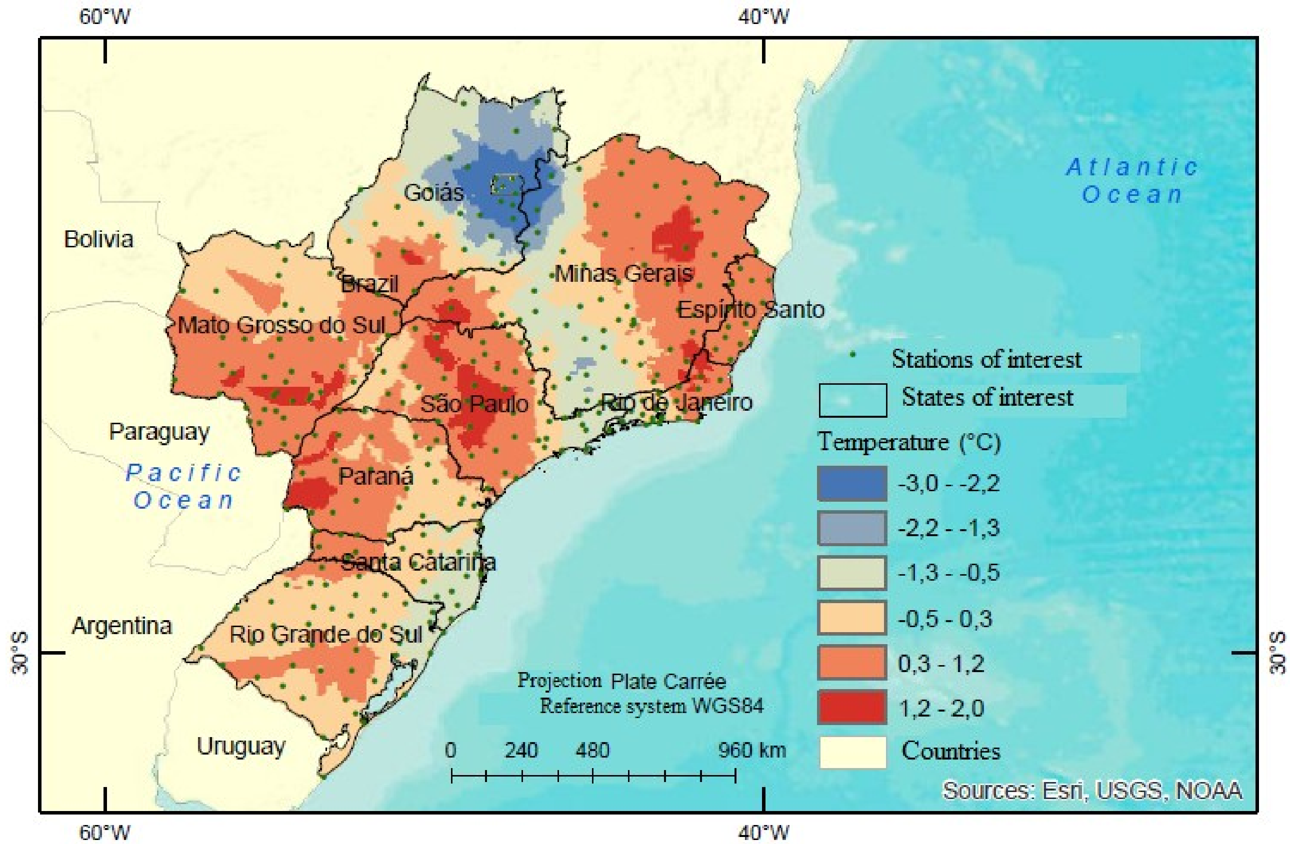

Figure 12 – Difference between observed and estimated temperatures.
Source: Authors (2023).

Figure 12 represents the difference between the temperature observed by the INMET stations and the temperature obtained in the model. Therefore, it can be inferred that the model underestimated the values in the reddish regions and overestimated the values in the bluish regions. However, it can be observed that the residuals ranged from -4.91°C to 5.35°C, but 67% were concentrated in the range between -1.86°C and 1.24°C. This suggests that the majority of the residuals fall within a small range, demonstrating good adequacy of the model. Understanding how air temperature varies spatially is essential for understanding and studying the climate of a particular region, establishing agroclimatic zones, assessing climate risks for agricultural and forestry activities, identifying drought and desertification events, analyzing the distribution of native plant species. Additionally, knowledge of the current spatial distribution patterns of air temperature is crucial for assessing the impacts of climate change. This information supports, for example, socio-environmental, credit, and insurance policies for rural and forestry activities, which can be useful for biodiversity conservation by promoting sustainable agricultural and forestry practices (ASSAD et al., 2020).

4. Final Considerations

In this study, temperatures for the study area were estimated for August 8, 2021, using the variables of altitude, latitude, longitude, and surface temperature. The obtained model showed efficiency in estimating air temperature. The equation obtained has a correlation index $R^2 = 0.7431$, which can be considered satisfactory. The p-values of the explanatory variables indicate that they can be used to estimate the value of the dependent variable.

The highest estimated temperatures are located in the Central-West region, demonstrating the influence of continentality in the environment. The lowest values are mostly present in regions of high altitudes, illustrating the influence of this variable on temperature.

However, the method has limitations, such as the scarcity of meteorological data in Brazil, which does not have an extensive network of meteorological stations, especially in the North and Northeast regions, which is why such regions could not be considered in the study.

It was also found that the highest residual values occurred in regions of high altitudes, or in regions where meteorological stations had shorter collection times during the studied day. Such situations make it difficult for the model to make an accurate estimate of the region's temperature.

Future research could be conducted by adding more environmental variables to the analysis model. NDVI (Normalized Difference Vegetation Index), precipitation, and land cover are examples of environmental variables that can be added to the analysis to improve the accuracy and understanding of air temperature behavior in the environment. Additionally, it is also suggested that models based on artificial neural networks be created to assess the accuracy of each methodology comparatively.

References

ASSAD, E. D.; VICTORIA, D. de C.; CUADRA, S. V.; PUGLIERO, V. S.; ZANETTI, M. R. Effect of Climate Change on Cerrado Agriculture. In: BOLFE, E. L.; SANO, E. E.; CAMPOS, S. K. Agricultural Dynamics in the Cerrado: Analysis and Projections. Brasília: Embrapa, 2020. p. 213–228.

Capuchinho, F. F.; Siqueira, M. P.; Cruz, G. H. T.; Costa, L. S.; Reis, E. F. Geographic Coordinates and Altitude in the Estimation of Air Temperature in the State of Goiás. Revista Brasileira de Agricultura Irrigada-RBAI, v. 13, n. 2, p. 3346-3357, 2019.

Paula, D. C. J., Guarienti, C. E., Gomes, R. de S. R., & Santos, F. M. de M. Estimation of Microclimatic Air Temperature in Cuiabá/MT using Neural Networks. Revista Brasileira De Climatologia, v.28, 184–200, 2021.

Filgueiras, R.; Nicolete, D. A. P.; Carvalho, T. M.; Cunha, A. R.; Zimback, C. R. L. Prediction of Air Temperature through Orbital Remote Sensing. Revista brasileira de cartografia, v. 68, n. 8, 2016.

Lapponi, J.C. Statistics Using Excel. 4th ed. Rio de Janeiro: CAMPUS-RJ, 2005, 476p.

Lima, M. T. V. Estimation of Average Air Temperature in ES. Equador, Teresina, v. 9, n. 2, p. 1-18, 2020.

Martins, A. P.; Alves, W. dos S.; Damasceno, C. E. Evaluation of interpolation methods for spatialization of air temperature data in the Paranaíba River basin – Brazil. Revista brasileira de climatologia, v. 25, p. 444-463, 2019.

Medeiros, S. DE. S.; Cecílio, R. A.; Melo Junior, J. C. F. DE.; Silva Junior, J. L. C. Estimation and spatialization of minimum, mean, and maximum air temperatures in the Northeast region of Brazil. In: Brazilian Journal of Agricultural and Environmental Engineering, Campina Grande, v.9, n.2, p.247-255, 2005.

Mendonça, F.; Oliveira, I.M. Climatology: Basic Concepts and Climates of Brazil. São Paulo: Oficina de Textos, 2007. 208p.

Montgomery, D.C.; Peck, E. A.; Vining, G. G. Introduction to Linear Regression Analysis. 5th ed. [S.l.]: John Wiley & Sons, 2012. 672p.

Nimer, E. Climatology of Brazil. Rio de Janeiro: IBGE, Department of Natural Resources and Environmental Studies, 2nd ed. 1989. 421p.

Nimer, E. A methodological model for climate classification. Brazilian Journal of Geography, Rio de Janeiro: IBGE, year 41, n. 4, p. 59-89, Oct./Dec. 1979.

WAN, Z. MODIS Land Surface Temperature Products Users' Guide. ERI, University of California, Santa Barbara, 2019. Available at: https://modis-land.gsfc.nasa.gov/pdf/MOD11_User_Guide_V61.pdf. Accessed 06/20/2023.