Use of HANTS algorithm in Sentinel-2 images for interpolation of missing data: a case-study for soy agricultural areas in periods of growth and harvesting

Leandro F. Coladello[1], Clodoaldo S. Faria Junior[1,2], Alisson F. C. do Carmo[1]

[1]Inspectral, Presidente Prudente, São Paulo 19060-900, Brasil – (Leandro.col, clodoaldo.souza,alisson)@inspectral.com.br
[2]Departamento de Cartografia, Universidade do Estado de São Paulo (UNESP) em Presidente Prudente, São Paulo 19060-900, Brasil – clodoaldo.souza@unesp.br

Keywords: HANTS, Time Series, Agriculture, Image Processing, Remote Sensing

Abstract:

Information about the current state of an agricultural culture can be fundamental in decision making, varying from quantity of fertilizers to be used in an area to estimating the productivity of a harvest in a period to establish credits payment. Satellite images with reasonable spatial resolution like Sentinel-2 are a powerful tool to visualize agricultural areas production, especially with the application of multispectral indexes like the Normalized Difference Vegetation Index (NDVI). Yet, due to the limitation of temporal resolution of satellites and periods with large cloudy weather can make the analysis difficult. This study proposed the use of the harmonic analysis time series algorithm (HANTS) to aid the incompleteness of time series obtained by the available satellite images in a case study of soy culture in Brazil. By applying the algorithm, a harmonic interpolation is obtained to produce a full daily NDVI time series from the beginning to end of the study period, facilitating further analysis to obtain metrics of interest. In this study, time series with gaps on specified periods are created based on full time series, then interpolated and compared with the real ones based on root mean squared error (RMSE) to access its accuracy.

1. Introduction

Vegetation monitoring is very important for agricultural areas since precise information of its current growth stage, from harvesting to planting, can be useful for fertilization management, pest control, and other harvest operations like estimative of productivity of the area or percentage of loss of a crop (Ye et al., 2023).

Most of vegetation monitoring methods involve the use of remote sensing imagery with the application of some index that use the red and near-infrared bands, like the Normalized Difference Vegetation Index (NDVI), which highlights areas with potential greenness of vegetation. However, due to the temporal resolution of satellites orbiting through the planet, the probability of the date of the image being taken by the satellite being contaminated by cloud interference, the non-possibility of images being taken during night by passive sensors, and the cost of high-resolution imagery, it is usual to not have a daily (or even an equidistant) collection of images for analysis of the crop development (Jia et al., 2021, Li et al., 2022).

Using collections of free satellite imagery available worldwide with the aid of interpolation methods to obtain information about the gaps caused by missing data due to the causes mentioned previously can be useful for crop monitoring. Dataspace Copernicus offers a plethora of image collections, like Sentinel-2 with a temporal resolution of nearly a week depending on the number of satellites revisits and a spatial resolution of 10m for the essential red and near-infrared bands for the calculation of NDVI (Phiri et al., 2020). With a reasonable small number of available images, harmonic analysis of time series (HANTS) can be used for interpolation of missing data, accounting for cloud contamination and outliers removing (Roerink et al., 2020). This study proposed the use of HANTS for interpolating missing data for crop growth analysis during epochs of planting and harvesting in a soy field case-study using NDVI as a proxis for vegetation assessment.

Soy has extremely importance in Brazilian harvests as is it the world’s largest producer, with an approximately 44-million-acre area being occupied by soy during 2022 and 2023 harvest, with an average productivity of 3508 kg/acre (EMPRAPA, 2024).

2. Background

Considering that an image can be seen as a numerical matrix, pixel information of an attribute of interest (like an index) can be localized at a position *(x, y)*. By extracting this information for each image at this position, a time series for that specific pixel is then obtained, as show in Figure 1 (Jönsson, 2004).

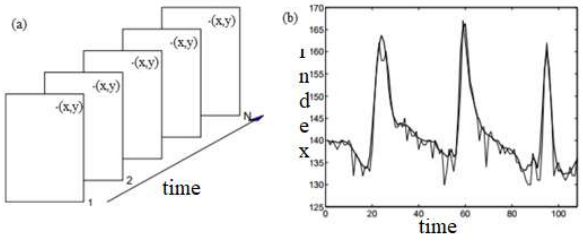


Figure 1. Pixel-wise image time series where (a) is a sequence of images in different periods (*t=1, …, N*) and (b) is the time series obtained by extracting the pixel value for all images at the position (*x, y*).

For this study, the attribute of interest is NDVI, which is a multispectral index that tries to measure the greenness and density of a vegetation in a satellite image by comparing the ratio of the visible read and near infrared bands, which have high spectral response for healthy vegetation, specially. The formula is given by:

$$NDVI = \frac{NIR-RED}{NIR+RED} \quad (1)$$

where NIR and RED are the reflectance for the near infrared (at 0.76-0.9μm) and red bands (at 0.63-0.69 μm), respectively. The index varies from -1 to 1, indicating high density vegetation when the value is near 1, water and clouds when the value is negative, and bare soil when the values are positive and near zero (Verstrate and Pinty, 1996).

Since satellite images are not always available due to factors like temporal resolution (Sentinel-2, for example, has a 5-day approximately temporal resolution) or weather circumstances like clouds, which can mask the true value of the information of interest, HANTS, a signal processing tool whose core is a Fourier transformation (Malamiri et al., 2020) was studied to address those issues. With the use of harmonic analysis, a full daily time series could be obtained for the entire period of study.

Firstly, any periodical function can be written in a Fourier form series:

$$y_t = a_o + \sum_{j=1}^{M} a_j \cos(w_j t_i - \phi_j) \quad (2)$$

where M is the number of Fourier series frequency, a_o is the average of all y observations ($i = 1, \dots, N$), ϕ_j and a_j are the phase and amplitude of the $j-th$ harmonic term, respectively, and w_j is the frequency of the $j-th$ term, and is given by:

$$w_i = \left(\frac{2\pi}{N}\right) i, i = 1, 2, \dots, N \quad (3)$$

Since soy culture behavior can be defined as a culture with cyclical behavior, with its cycle during an average of 118 days (Andrade et al., 2019), HANTS could be applied. Figure 2 shows the cycle of soy culture with its respective months and stages of growth (according to Brazil’s agriculture calendar), being October the start stage with the crop emergence until February when the crop reaches full maturity, usually.

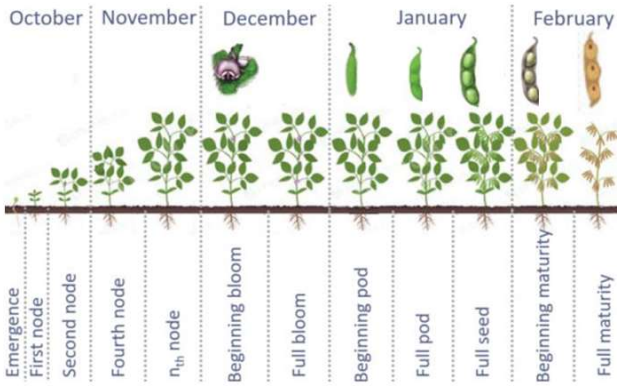


Figure 2. Soy cycle according to usual months and stages of growth from emergence (October) to full maturity (February). Source: Adapted from Vieira (2020).

Given known w_j and M, ϕ_j and a_j were the parameters for the HANTS algorithm, obtained by fitting the observations of the time series. To get a reliable signal recreated, several other hyperparameters must be set by the user, which are: hi/lo suspension flag, invalid data rejection threshold, fit error tolerance and degree of over determinedness. These parameters were well described in Roerink et al. (2000) and in Table 1 the most common settings for them were show (Zhou et al., 2016).

Parameter	Setting
High range of parameter	1
Low range of parameter	-0.2
High/Low Flag	Low
Degree of overdetermination	5
Number of frequencies	3
Base period	365
Fit error tolerance	0.05
Regularization factor	0.5

Table 1. Parameters of HANTS and their mostly used settings in NDVI studies. Source: Adapted from (Zhou et al., 2016).

The first two parameters of Table 1 indicate the range of parameter in analysis, and if exceeded, they are considered outliers and are not considered in further analysis. The flag parameter indicates if outliers are expected above or below the fitted model and since cloud contaminated pixels or bad productivity produces low NDVI values, it is usual to set it to “low”. The degree of overdetermination is the number of extra valid observations that can be inputted in the model. The minimum number of observations required to fit the curve is given by

$$2 \times number\ of\ frequencies - 1 \quad (4)$$

Then, the number of frequencies parameter is defined as the number of harmonics (excluding zero-frequency) to be used in the reconstruction model and, as a rule of thumb, the number is usually set to the estimated periodicity of the harvest plus three units. For NDVI with an annual peak, for example, the number of frequencies can be at least four. If NDVI shows two peaks, then at least 5 frequencies should be used. The base period is usually set to 365 in remote sensing-based NDVI data as it dominated by seasonal and yearly variations. Fit error tolerance is the maximum deviation between the real observations and the results obtained by the reconstruction and finally the regularization factor is usually set to a small value, usually from 0.1 to 0.5 to control high amplitudes of the model (Roerink et al., 2020, Zhou et. al, 2021).

3. Materials and methods

The methodology for this study is presented in the flowchart of Figure 3. Sentinel-2 images including area of soy fields located in Brazil were downloaded via *Google Earth Engine* (GEE) for the years of 2019 and 2020. Download of images, all processing steps and analysis results were achieved by using the *Python* language using well known libraries like *numpy, pandas* and *GDAL*.

After downloading the images, only the area representing soy fields were cropped from the full images for the presented study. Sentinel-2 images from GEE have atmospheric, terrain and cirrus correction via *sen2cor* algorithm (ESA, 2022), thus were ready to be used for NDVI analysis via equation (1). RED and NIR bands have 10 meters of spatial resolution, so resampling was not needed.

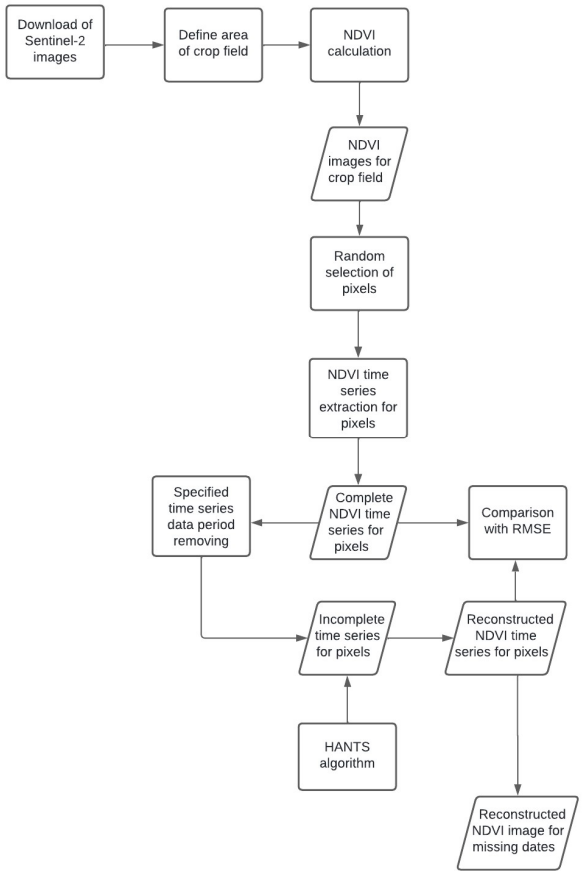


Figure 3. Methodology applied in this study.

From the collection of NDVI images, a random sample of pixels were selected for the soy fields, and their temporal behavior were plotted for inspection. The results were shown in Figure 4. From the known time series profile of the selected pixels, four delimited time stamps were selected to be removed from the time series and identified in red rectangles (A, B, C and D). These periods indicate cycles of harvesting (A and B) and growing (C and D) and is in accordance with the periods shown in Figure 2, i.e., it can be observed a new cycle starting near period C (October) with emergence of new crops and maturity being reached near periods A and D, approximately in February.

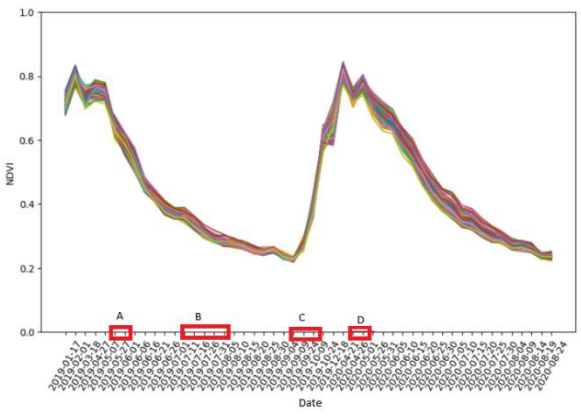


Figure 4. NDVI profile for randomly selected pixels from soy crops identified pixels.

In Figure 5 it is presented examples for (i) harvesting periods (like A in Figure 2) and (ii) growing periods (like C in Figure 2) of (a) Sentinel-2 images fragments where soy fields were identified, in RGB composition, (b) the cropped soy field in RGB composition, and (c) NDVI calculated for these areas.

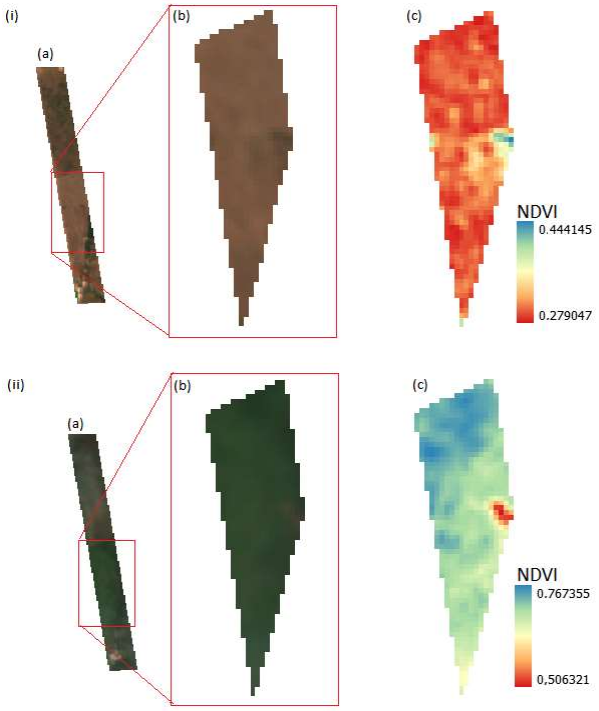


Figure 5. (i) Sentinel-2 image strip (a) where a soy field is in a harvesting period (b), in which NDVI was calculated (c). (ii) Sentinel-2 image strip (a) where a soy field is in a growth period (b), in which NVDVI was calculated (c).

Then, HANTS algorithm was applied to all random sampled pixels time series with the four gaps (A, B, C, and D) to reconstruct the now unknown values of NDVI for these gaps and to study the behavior of the reconstruction. For this, root mean square error (RMSE) and symmetric mean absolute percentage error (sMAPE) were used to evaluate the performance and accuracy of the algorithm for this case study. The RSME and sMAPE formula are given by:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}} \quad (4)$$

$$sMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \quad (5)$$

with y_i being the reference NDVI value, $\hat{y}_i$ the predicted NDVI value by the HANTS algorithm and n the number of pixels in the image. sMAPE was preferred over mean absolute percentage error due to many values near zero, which can be problematic when computing MAPE values.

Finally, an estimated image for the entire area can be created by getting each time series value for all pixels in a specified date. The Nearest Neighbor method (Taunk et al., 2019) was used as a

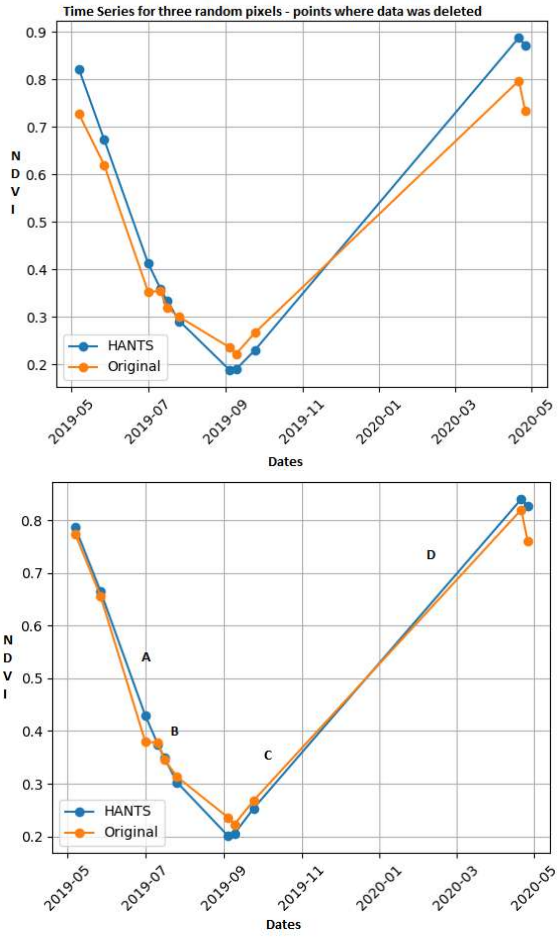
resampling algorithm to create the pixel visual representation in the same spatial resolution as the original images.

4. Results and discussion

By calculating NDVI from equation (1) for the area of interest, it can be observed high values like the ones from Figure 5 when soy is in growing period (above or equal 0.5 in general), and low values (below 0.5 in general) when crop is in its maturity stage and/or the harvest period is occurring.

The selected area was represented in 22×69 (latitude, longitude) grid and then $22 \times 69 = 1518$ time series were estimated with HANTS.

Figure 6 shows the reconstructed NDVI by using HANTS for the gaps A, B, C and D for a time series of three single pixels. By visual inspection it is possible to observe that HANTS could maintain the behavior of the original time series with reasonable small error. Then, this analysis was replicated for every pixel of the area and the RMSE and sMAPE for each one was inspected, as shown in Figure 7 and 8.



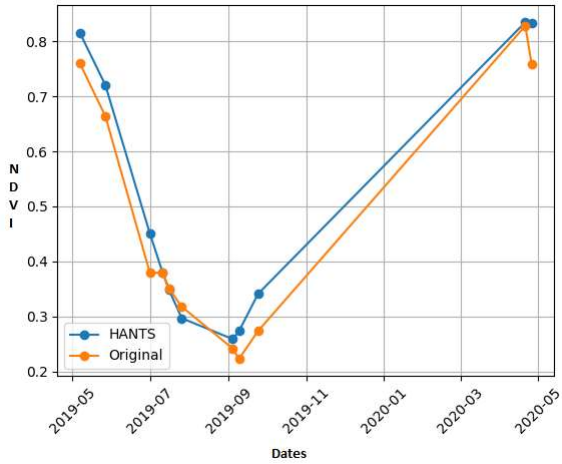


Figure 6. Reconstructed NDVI by HANTS for the deleted gaps (A, B, C and D) in the time series for three random pixels and their original values.

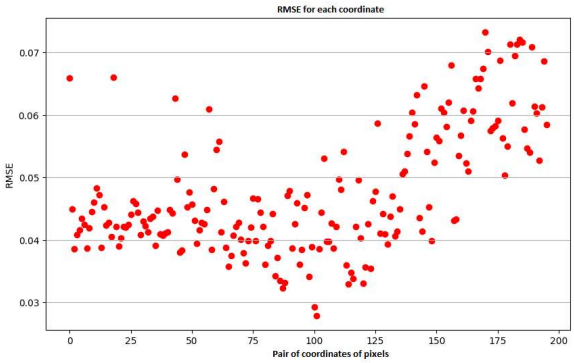


Figure 7. RMSE for reconstructed NDVI for every time series pixel in the studied area (total of 196 time series).

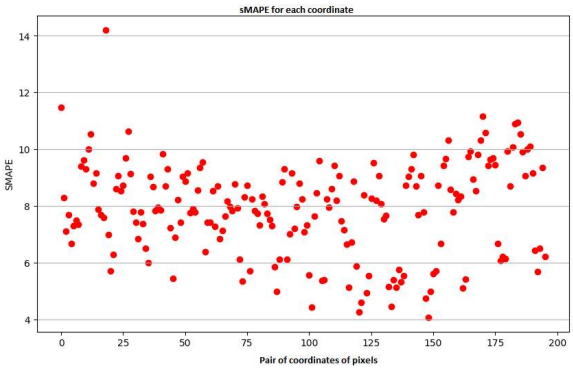


Figure 8. sMAPE for reconstructed NDVI for every time series pixel (total of 196 time series).

Table 2 shows the mean of each metric for all pixels. The maximum RMSE value obtained was approximately 0.08 while the maximum sMAPE value obtained was approximately 14%, while most of errors were found below 12% error indicating a good HANTS performance for the area studied.

Metric	Mean
RSME	0.047580
sMAPE	7.88335%

Table 2. Mean of accuracy metrics for studied area.

By getting the time series values for all pixels of a missing date, it was possible to create a visual representation for the entire area during the absence of data, as shown in Figure 9.

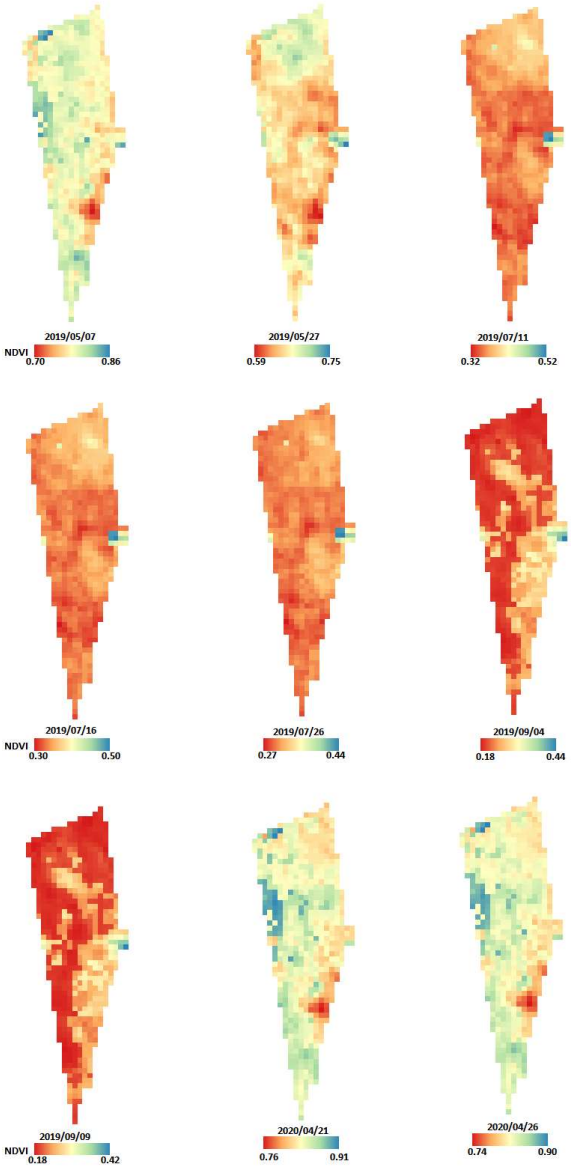


Figure 9. Estimated NDVI images from HANTS for missing dates during the study period.

As can be seen in the estimated images, they capture the pattern presented in the pixel only time series of Figures 4 and 6, where there is a decrease of NDVI during the end of July to September of 2019, where NDVI reaches its bottom value for the entire study period (which is highlighted by the reddish tones in those images), and then, by April of 2020, NDVI becomes higher again, varying from 0.75 to 0.91.

5. Conclusion

Brazil is the main producer of soy in the world and so, areas representing soy culture were selected to test HANTS algorithm. This algorithm is a harmonic analysis for time series which performs well when cyclical behavior is expected. Since soy culture has this kind of behavior, mainly from October to February in Brazil, it was possible to show that HANTS could interpolate well missing data during different stages of the growth cycle of soy in this case study. Accuracy metrics were gotten as it could be seen from small RMSE (<0.07 in most cases) and small sMAPE (below 15%). Estimated images for all missing dates could also be generated using HANTS. The full complete time series facilitates decision making regarding estimative of production through the entire cycle, estimative of loss of harvesting, total area harvested and others.

References

Andrade, D. M., Turatti, P., Moura, V., Souza, R. A. 2019. *Comportamento espectral do ciclo fenológico da soja com uso do EVI MODIS/MAIAC e EVI MOD13A2. Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto*.

Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA). 2024. *Soja*. Available at: https://www.embrapa.com.br/soja. Last visited on July 07th.

European Space Agency (ESA). 2024. *Sen2cor 2.11 configuration and user manual.* Available at: https://step.esa.int/thirdparties/sen2cor/2.11.0/docs/OMPC.TPZG.SUM.001%20-%20i1r0%20-%20Sen2Cor%202.11.00%20Configuration%20and%20User%20Manual.pdf. Last visited on July 6th.

Jia, J., Sun, H., Jiang, C., Karila, K., Karjalainen, M. et al. 2021. *Review of active and passive remote sensing techniques for road extraction. Remote Sensing*, v. 13 (21).

Jönsson, P.; Eklundh L. 2004. TIMESAT - a program for analyzing time-series of satellite sensor data. Computers & Geosciences, v. 30, n. 8, p. 833 – 845.

Li, Z., Shen, H., Weng, Q., Zhang, Y., Dou, P., Zhang L. 2022. *Cloud and cloud shadow detection for optical satellite imagery: Features, algorithms and prospects. ISPRS Journal of Photogrammetry and Remote Sensing*, v. 188.

Malamiri, H. R. G., Zare, H., Rousta, I., Olafsson, H. et al. 2020. *Comparison of harmonic analysis of time series (HANTS) and multi-singular spectrum analysis (M-SSA) in reconstruction of long-gap missing data in NDVI Time series*, v. *12 (17).*

Phiri, D., Simwand, M., Salekin, S., Nyrienda, V., Murayama, Y., Ranagalage, M. 2020. *Sentinel-2 data for land cover/use mapping: a review. Remote Sensing*, v. 12 (14).

Taunk, K., De, S., Verma, S., Swetapadma, A. 2020. *A Briefing review of nearest neighbor algorithm for learning and classification. IEEE International Conference on Intelligent Computing and Control Systems (ICCS).*

Roerink, G., J., Menenti, M., Verhoef, W. 2000. *Reconstructing cloudfree NDVI composites using Fourier Analysis of time series. International Journal of Remote Sensing,* v. 21 (9).

Verstraete, M. M., Pinty, B. 1996. *Designing optimal spectral indexes for remote sensing applications. IEEE Transactions on Geoscience and Remote sensing, v. 34 (5).*

Vieira, V. 2024. *Florescimento Soja*. *Instituto Federal do Mato Grosso do Sul. Ministério da Educação.* Available in: https://www.ifms.edu.br/imagens/imagens-noticias/materia-soja. Last visited on July 08th.

Ye, J., Bao, W., Liao, C., Chen. D., Hu, H. 2023. *Corn phenology detection using the derivative dynamic time warping method and Sentinel-2 time series*. *Remote Sensing*, v.15, 3456.

Zhou, J., Jia, L., Hoek, M., Menenti, M., Hu, J. L. 2016. *An optimization of parameter settings in HANTS for global NDVI time series reconstruction. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*.

Zhou, J., Jia, L., Menenti, M., Liu, X. 2021. *Optimal estimate of Global Biome – specific parameter settings to reconstruct NDVI time series with the harmonic analysis of time series (HANTS) method. Remote Sensing,* v. 13, 4251.