A COMPARATIVE STUDY OF TIME-SERIES RECONSTRUCTRION METHODOLOGIES ON THE GOOGLE EARTH ENGINE PLATFORM

Randy Chua[1]*, Chong Chen[1], Bin Zhou[1], Huizhen Zhou[2], Jianjun He[2]
[1] *Twenty First Century Aerospace Technology (Asia) Pte. Ltd., 61 Science Park Road, #05-17 The Galen, Singapore Science Park II, Singapore 117525*
[2] *Twenty First Century Aerospace Technology Co., Ltd., No. 26 Jiancaicheng East Road, Haidian District, Beijing 100096, China*
Email: randy.chua@21at.sg; chenchong@21at.sg; zhoubin@21at.sg; zhouhz@21at.com.cn; hejj@21at.com.cn

KEY WORDS: Time-series Reconstruction, Classification, Google Earth Engine, Sentinel-2, TripleSat

ABSTRACT: Consistent long-term time-series information derived from satellite imagery is especially important in the fields of vegetation-related research. This ensures that phases of vegetative growth or changes can be adequately captured for various agricultural or environmental applications, such as species identification, yield predictions, and biomass management. However, commonly used optical satellite data, like the Landsat and Sentinel-2, are susceptible to contamination by the presence of clouds, aerosols, and other noise sources. Derived time-series data under these conditions may be temporally inconsistent due to invalid or missing temporal points, thereby reducing the effectiveness of high temporal resolutions boasted by these readily available remote sensing data catalogues. To overcome these limitations, we adapted the temporal smoothing concept of incorporating temporal sliding windows and generation of weighted regression lines along time-series information, and proposed three different reconstruction methods: the Local-Local reconstruction centers temporal windows on existing temporal points, and reconstruct values in-situ; Interval-Local creates temporal windows at fixed intervals, and recalculates values of existing points; Interval-Interval creates window intervals, but recreates temporal points at newly-defined interval sections.

The time-series reconstruction methods are implemented on the Google Earth Engine (GEE). GEE provides a huge database for accessing updated satellite imagery, and a cloud platform to process and compute huge volumes of satellite data. The different reconstruction methods are applied in the classification of Mumbai City's vegetation classes using Sentinel-2 Level-2A imagery. TripleSat high-resolution optical imagery is used to identify different forms of vegetation parcels for verification purposes.

The effectiveness of each method are evaluated based on their performance in accurately reconstructing time-series patterns and identifying characteristics of the associated vegetation type within the study area. While the L-L method appears to be the most computationally efficient on the GEE platform, and is capable of preserving the temporal integrity of the input Sentinel satellite series data; the I-I method, in reconstructing temporal patterns with more regular intervals, resolves temporal inconsistency issues due to missing data points. The regular sampling and recalibration of time-series data points at intervals allow for a more realistic reconstruction of the temporal patterns that may not have been present in less consistent time-series datasets. This allows for reconstructed time-series information to be better optimized for, and serve as a stronger foundation in future vegetation or biophysical studies, where complex and consistent temporal information will be essential and appreciated.

1. INTRODUCTION

Consistent time-series information derived from remotely sensed satellite imagery products have been a staple input into the studies of the Earth's myriad biophysical surfaces. Temporal patterns and variations of spectral reflectance provided by optical satellites such as the Sentinel-2 and the Landsat constellations have assisted researchers in the identification, differentiations, and characterizations of the Earth's land cover types.

However, optical satellite imagery is susceptible to numerous problematic interferences, such as clouds and aerosol contamination obscuring optical observations; or instrumental calibrations or failures causing derived time-series information to be temporally and spatially incomplete (Goward et al, 1991). Missing information affects accuracy of their temporal characteristics, and limits the range of their applicability. It is essential to address this phenomenon, to improve time-series data's applications towards remote sensing studies.

Studies have categorized methodologies towards reconstructing these missing information of values from remote sensing products into three main categories: spatial-based, spectral-based, and temporal-based methods. Spatial-based methods are some of the commonly used approaches, involving correlating missing data with locally available pixels, and applying spatial correlation algorithms to restore the missing pixels. Zhang et al (2007) adopted the kriging geostatistical approach to resolve the SLC strip error in the Landsat 7 ETM product, while Xu et al (2015) re-computed missing MODIS NDVI data using the linear interpolation algorithm. Spatial-based approaches often carry an inherent assumption that the missing pixels, and those that are used in the interpolation process, share similar geographical and thus spectral structures (Guillemot & Le Meur, 2014).

Spectral-based approaches involve the reconstruction of a single spectral band by correlating with complementary spectral bands within similar satellite products. This solution is based on the assumption that the missing region is geographically different in the various spectral bands (Shen et al, 2015), and the spectral characteristics can be adequately captured by other bands. Multiple studies have been done to retrieve the faulty Aqua MODIS band 6 that had arisen due to defective detectors, by studying its relationship with both the Aqua band 7, and the Terra MODIS spectral bands (Wang et al, 2006).

Temporal-based methods, on the other hand, derive values for missing pixels from the temporal spectrum, by extracting relevant information from other satellite imagery acquired at the same region, but at different time periods. Missing data due to cloud cover, for example, may affect regions too large for spatial interpolations to be reliable. It can also affect spectral information collected across most, if not all, spectral band wavelengths. However, the short-term interference on data collection caused by such atmospheric conditions can usually be mitigated by extending into multi-temporal datasets, whereby spectral information of a single missing pixel can be interpolated from available data points within temporal windows centered around the missing temporal point. Shen et al (2015) pointed out that temporal-based approaches are not suitable in retrieving spectral changes caused by abrupt geographical transformations, as temporal methods cannot reintroduce new or non-existing patterns not present in the time series. However, they are adequate in capturing and reconstructing the spectral characteristics of most geographical features with suitable time intervals being considered.

Common temporal-based methods include fitting algorithmic functions such as the Harmonic Analysis of Time Series (HANTS) into time series data (Jakubauskas et al, 2001; Zhou et al, 2015); temporal filter applications, such as the Savitzky-Golay filter (Chen et al, 2004; Zhou et al, 2016); reconstruction via deep-learning models (Das & Ghosh, 2017); and temporal interpolation methods (Julien & Sobrino, 2010). Function fitting approaches such as the HANTS require the prior assumption that time series information already adheres to certain temporal patterns, which may not be the case in most natural environmental conditions. Deep-learning approaches, on the other hand, often require huge amounts of input data and computational capabilities, and face issues such as resolution inconsistencies. A weighted reconstruction method was proposed by many scholars to counteract the contamination of remote sensing-derived signals by atmospheric conditions during the interpolating or filtering processes (Swets et al, 1999).

This paper proposes three different time series reconstruction methods, incorporating concepts of temporal interpolation, filtering, and weighted regressions. Two parameters are differentiated between the methods: the nature of moving temporal windows, and the temporal intervals of the reconstructed time series. These methods are implemented on the Google Earth Engine (GEE) platform. A vegetation classification case study in Mumbai City during the Rabi season is used to evaluate the proposed methods. The Normalized Difference Vegetation Index (NDVI), a commonly-used remote sensing indicator, is used in this case study as the target for reconstruction. Sentinel-2 optical imagery datasets are used as the primary time series data for this paper, and a high-resolution imagery source, TripleSat, is used for ground-truth validation purposes.

The remaining sections of the paper are as structured: Section 2 describes the study area, datasets, and platform used in this paper. Section 3 details the methodologies of the three proposed reconstruction approaches, parameters used, and the implementation processes. Section 4 covers the results and discussions of the classification and time series reconstruction results using the different methods. Section 5 concludes the paper by reviewing the advantages and disadvantages of the proposed methods, in both the contexts of the case study, and other real world applications.

2. STUDY AREA & DATASETS

2.1. Study Area

Mumbai City, a major urban district in Maharashtra, India, was selected as the test case study area in this paper to carry out vegetation type classification (Figure 1). Mumbai City serves as the state capital, and many different forms of vegetative land cover types can be found in the area. Outside of urban vegetation such as parks and open grasslands, the city also boasts areas of cultivated croplands, pastures, and fruit plantations. Forests can commonly be found within the Sanjay Gandhi National Park; and mangroves are very common in the region as Mumbai City as a result of land reclamation.

Most of the Indian continent, especially Mumbai City, experiences consistent cloud cover throughout the year, during both dry and wet seasons. As a result, most optical satellite imagery data of the region suffers from a certain degree of cloud contamination. Figure 2 depicts the extent of missing temporal information as a result of cloud cover during Mumbai City's 2019-2020 Rabi season. In order to properly observe vegetation patterns and changes within the city, there is a need to properly fill in the data gaps in between the temporal datasets to provide consistent optical remote sensing observations.

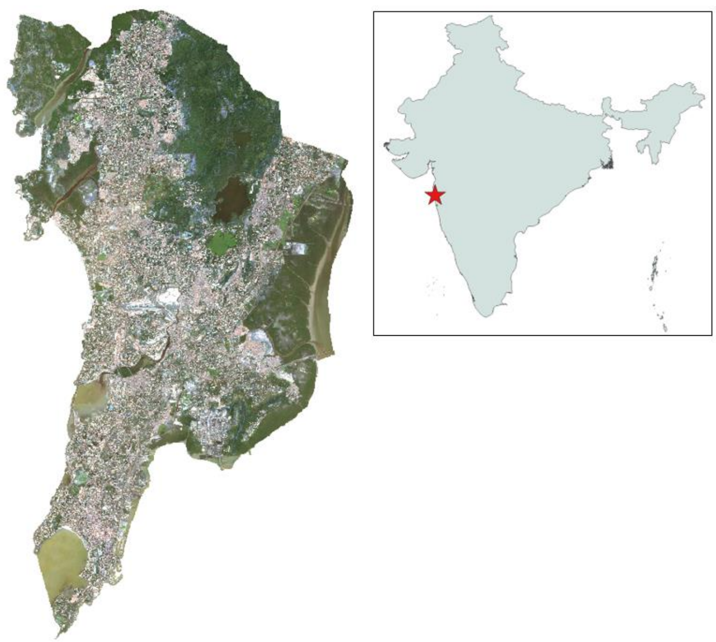

Figure 1. Location of the study region in Maharashtra, India.

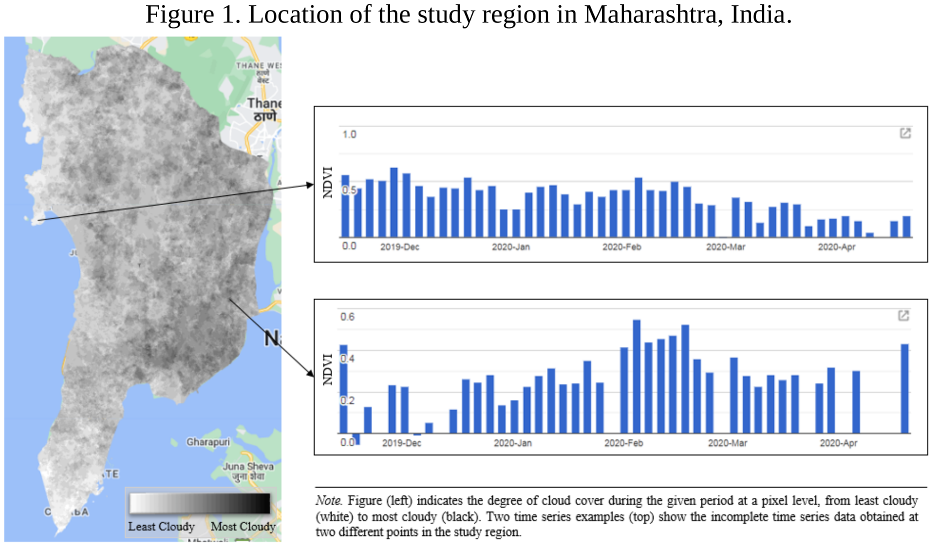

Figure 2. Extent of cloud contamination and affected time series in Mumbai City.

2.2. Datasets

Two sets of optical satellite imagery data are employed in this case study: Sentinel-2 and TripleSat constellations.

The Copernicus Sentinel-2 Level 2A products provide optical satellite imagery that are highly suitable for time series related remote sensing studies. The Sentinel-2 imagery boasts a spatial resolution of 10 meters; contains multiple spectral bands along the red edges and infrared red electromagnetic spectrum, which are highly sensitive to vegetation. Also, the Sentinel-2 has a global revisit rate of 5 to 7 days across its dual identical satellites, allowing frequent temporal data to be collected, and provide greater flexibility in selecting temporal filtering windows and intervals during reconstruction processes. NDVI reconstruction is chosen as the main focus for the subsequent methodologies, due to its relevance in most vegetation studies.

The high-resolution TripleSat satellite constellation is used as the verification dataset, due to it being capable of providing quality images with spatial resolutions up to 0.8 meters. Two TripleSat images are obtained for the study region, in the months of January and April 2020. These two images provide information on the vegetation conditions during the early and late stages of the dry Rabi season, and are suitable as visual verification datasets for the vegetation classification results.

2.3. Google Earth Engine

The GEE platform, on which the proposed methods are developed on, is a cloud-based platform that is easily accessible on various browsers and computer systems. GEE provides a Javascript/Python interface that allows users to both gain access to huge volumes of analysis-ready remote sensing datasets, and also provide high-performance parallel computing capabilities to carry out additional processing and applications (Gorelick et al, 2017).

As of date, the GEE platform provides access to historical and updated Sentinel-2 pre-processed imagery. GEE’s server-side cloud computing capabilities allow for huge temporal datasets of very specific geographical extents to be easily extracted and compiled, in the case of Mumbai City that spans two different Sentinel-2 tiles. GEE also computes algorithms and calculations on its own server, before displaying client-side, reducing the computational and graphical requirements on the user end, especially for the proposed methods that involve iterative sliding windows and regression calculations across multiple temporal images.

3. METHODOLOGY

3.1. Parameters

The general methodology revolves around the reconstruction of temporal data by generating weighted linear regression lines within moving windows along the input time series data, resulting in a group of regression lines and values associated with each predetermined temporal point (Figure 3).
Two different parameters constituting the reconstruction methods are proposed. The first parameter, the window shift, indicates how the temporal window used to select temporal points for interpolation shifts along the time series. The temporal window (yellow rectangle) can be either centered around local existing points (in blue), or alongside fixed user-defined intervals. The results are dependent on the availability and proximity of existing data points: centering around existing data points generates regressions and corresponding reconstructed values tailored to these exact points; however, in time series with irregularly spaced or sparse temporal points, the temporal windows and interpolations generated tend to become uneven and unequal across the entire time series.

The second parameter, the value reconstruction, refers to the temporal location of the value(s) in the reconstructed time series, which can either be a replacement of the existing local values, or at user-defined temporal intervals. Similar to the concept of the temporal windows, reconstructing time series at local temporal points preserve the temporal integrity of the underlying time series data, such as the standard twelve-day revisit rates of Landsat satellites; on the other hand, constructing new values from existing temporal data at user-defined intervals allow a greater control over the temporal structure of the reconstructed time series.

3.2. Reconstruction Approaches

Three different approaches are proposed in this paper: the Local-Local (L-L) reconstruction, the Interval-Local (I-L) reconstruction, and the Interval-Interval (I-I) reconstruction methods.

The first approach, the L-L reconstruction method, adopts local temporal windows and value reconstruction at individual existing local data points (Figure 4). A temporal window is constructed around each existing temporal point, and linear regression performed using all points within each window. The final value of the point is derived from the final regression line within the window, and replaces either the existing or empty value of the point. The L-L approach serves as a foundation method for the subsequent methods, due to its simplicity.

The second approach, the Interval-Local (I-L) reconstruction method, utilizes temporal windows at user-defined intervals, while maintaining value reconstruction at individual existing local data points (Figure 5).

Unlike the first approach, the temporal windows are centered around regular intervals (in red) along the time series. A new value is derived from the regression line calculated within each window. A new value is thus associated with each individual point, and this process continues as the temporal window moves down each interval along the time series. The result is that each existing point would have a family of regression lines and corresponding values associated with it. Cloud and other atmospheric influences have the tendency to reduce remote sensing data and associated vegetation indices such as the NDVI (Julien & Sobrino, 2010; Cao et al, 2018) A weighted average towards peak values for each point is proposed, to reduce the impact of these data interferences (Swets et al, 1999).

Lastly, the third approach, the Interval-Interval (I-I) reconstruction method, deploys user-defined intervals in both the shifting of temporal windows, and also in the construction of new values during the time series reconstruction process (Figure 6). Regular intervals (in red) are first constructed across the entire time series. At each interval, the existing temporal points are used in the linear regression calculations. However, unlike the previous two approaches, the corresponding values are instead derived for the interval points, rather than the existing local points. As the temporal windows slide down along the intervals, each interval point receives new regression values, and these values are similarly averaged with a weighted approach to produce a reconstructed time series at user-defined, regular temporal intervals.

3.3. Implementation

The three methods are implemented on GEE web-based Javascript Code Editor, to employ the data accessibility, and computational capability of the platform. 190 pre-processed Sentinel-2 images obtained across 47 unique dates within the Rabi season (between November 2019 to May 2020) are selected from the GEE catalog; of which, 21 (11%) of these images have cloud percentages over 50%, and an additional 25 images (17%) experienced more than 20% cloud cover. Each method is implemented individually on this temporal dataset. For the reconstruction approaches, both the temporal windows and intervals are set at 10-day to optimize the advantages of Sentinel-2's high revisit rate.

Three main land cover types are identified. The Tree class is defined as the highly vegetated regions in the study area, and covers forests and wetland mangroves in Mumbai City. The Grassland class refers to the lightly vegetated land covers, such as the grazing pastures, shrubland, and croplands. Due to the time period chosen for this paper, the cropland class is considered under the Grassland category due to low agricultural intensities during the selected study period. Lastly, the Others class covers regions with little to no vegetation activity, primarily the urban and bare land regions.

Figure 7 summarizes the workflow for Mumbai City's vegetation classification. A simple threshold classification method is chosen, using a simple NDVI vegetation ratio. An NDVI value larger than 0.3 is selected as the threshold for this paper. The vegetation ratio calculates the ratio of NDVI values larger than the threshold to all available NDVI values within a single time series pattern for each temporal point. The value of the vegetation ratio reflects the percentage of input satellite imagery agreeing that the specific point is a vegetated pixel. A maximum NDVI value threshold is used to further differentiate between tree regions and grasslands with high vegetation temporal intensity but much lower maximum vegetation growth.

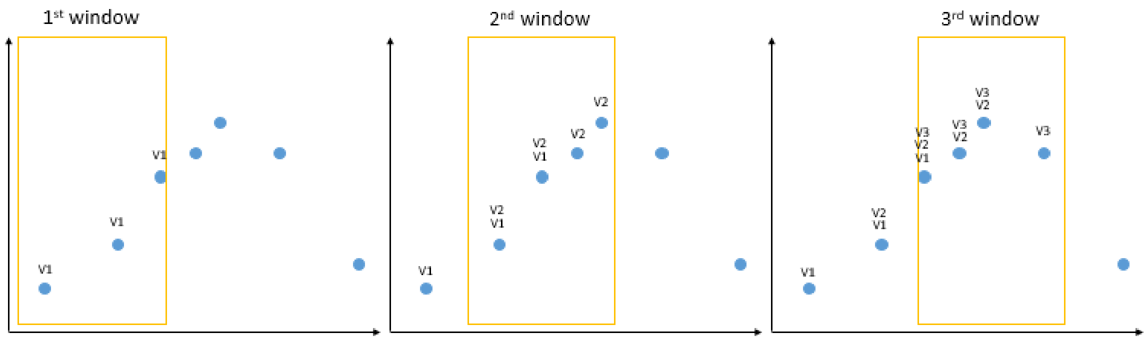

Figure 3. Illustration of moving temporal windows and associated values for each individual point.

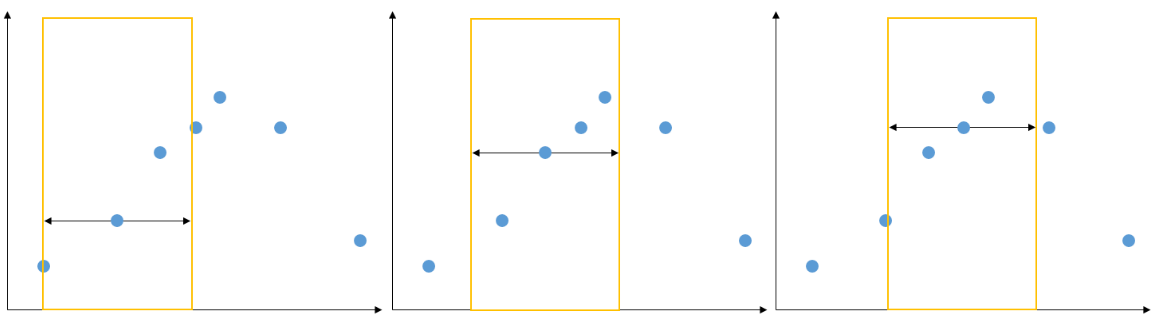

Figure 4. Illustration of a Local-Local reconstruction approach.

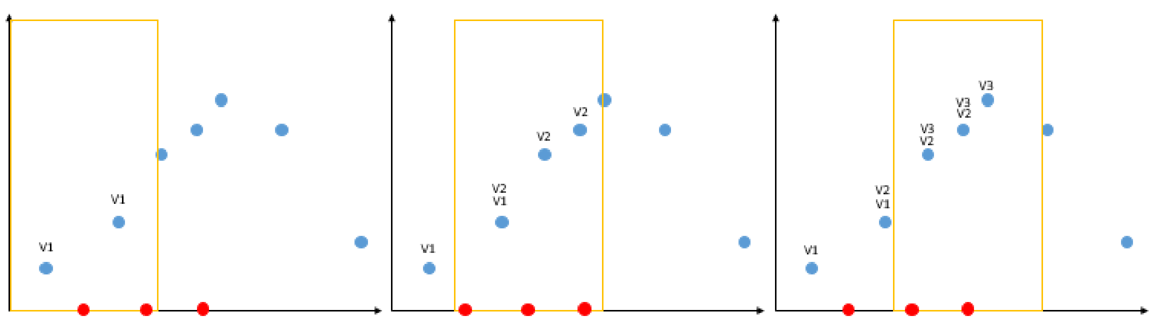

Figure 5. Illustration of an Interval-Local reconstruction approach.

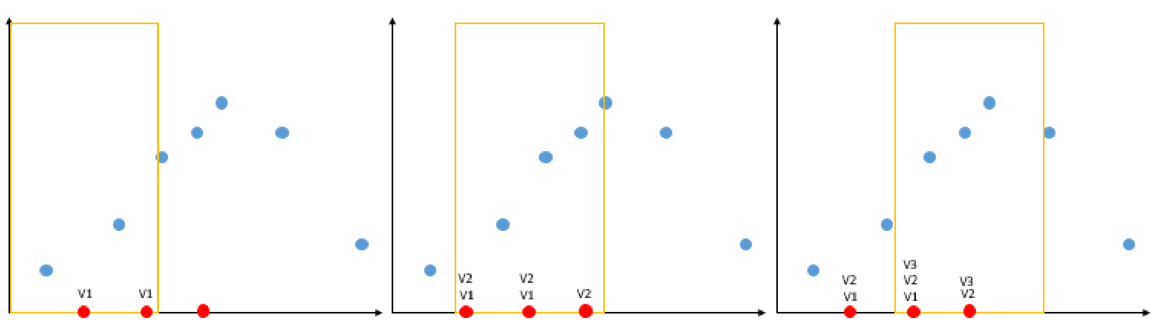

Figure 6. Illustration of an Interval-Interval reconstruction approach.

Table 1. Classification accuracy results using the original and reconstructed time series datasets.

	Original	Local-Local	Interval-Local	Interval-Interval
Accuracy (%)	71.3	87.1	86.0	86.4

4. RESULTS & DISCUSSION

Figure 8 shows some of the results of the classification using reconstructed Sentinel-2 time series datasets, and Table 1 shows the classification accuracies for each classification approach. Four sample regions are selected from the study region, and the results of each time series reconstruction method are compared against each other, and against the original time series pattern (Figure 9).

Time series patterns for data before February 2020 in Figure 9(A) demonstrates the consistency of time series reconstruction of the proposed methods against a well-defined and complete time series data: all three reconstructed time series follow the original temporal pattern within huge variations.

Differences between each method occur when the original time series patterns become irregular, either due to missing values, or anomalies. In Figures 9(B) and 9(C), the absence of a weighted average using multiple regressions with neighboring values (in the L-L method highlighted in red) cause abnormal values from the original time series to drag down the reconstructed values, resulting in an abnormal dip in signal during January 2020. The application of weighted averages towards higher values using multiple regressions across different temporal windows (in the I-L and I-I methods) allow for reconstructed time series to avoid abnormal dips in the original time series, and maintain a more consistent temporal pattern.

Extreme variations seen in the original time series in Figure 9(D) cause the L-L's method straightforward reconstruction method to return a nearly flat time series pattern that ignores potentially true land cover variations on the ground. Multiple regression averages by the I-L and I-I methods produce a smoother transition between these large value changes, producing eventual time series patterns that reflect possible large signal variations, while concurrently reducing extremities.

Between the I-L and I-I methods, while both reconstruction methods are able to preserve temporal patterns better than the L-L counterpart due to the weighted average approach, Figure 9 shows that the I-I method generally produces smoother time series patterns as compared to the I-L method. Compared to the inconsistency of the input time series intervals, such as the Sentinel-2's 5 to 7 day revisit rate while assuming no imagery issues, I-I method's reconstructed 10-day interval approach produces a resultant time series with strict regular temporal pacing, and is not bounded by the temporal variability of the input time series patterns and intervals. Introducing reconstructed values along regular temporal intervals using the I-I approach produces more consistent patterns as large temporal gaps between the input time series can be accounted for accurately by performing regressions and weighted averages at these specific temporal points. Working with time series data with regular intervals is beneficial for time-sensitive remote sensing studies, such as crop monitoring or phenological studies. Providing insights on values at required intervals or specific temporal stages are important when monitoring vegetative growth, and the I-I method provides the flexibility in the reconstructed time series.

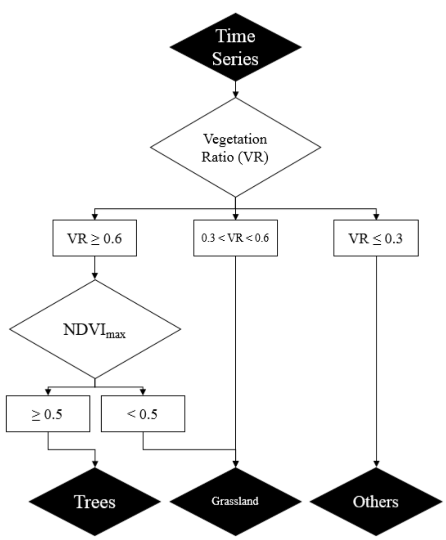

Figure 7. Simplified classification workflow using reconstructed time series information.

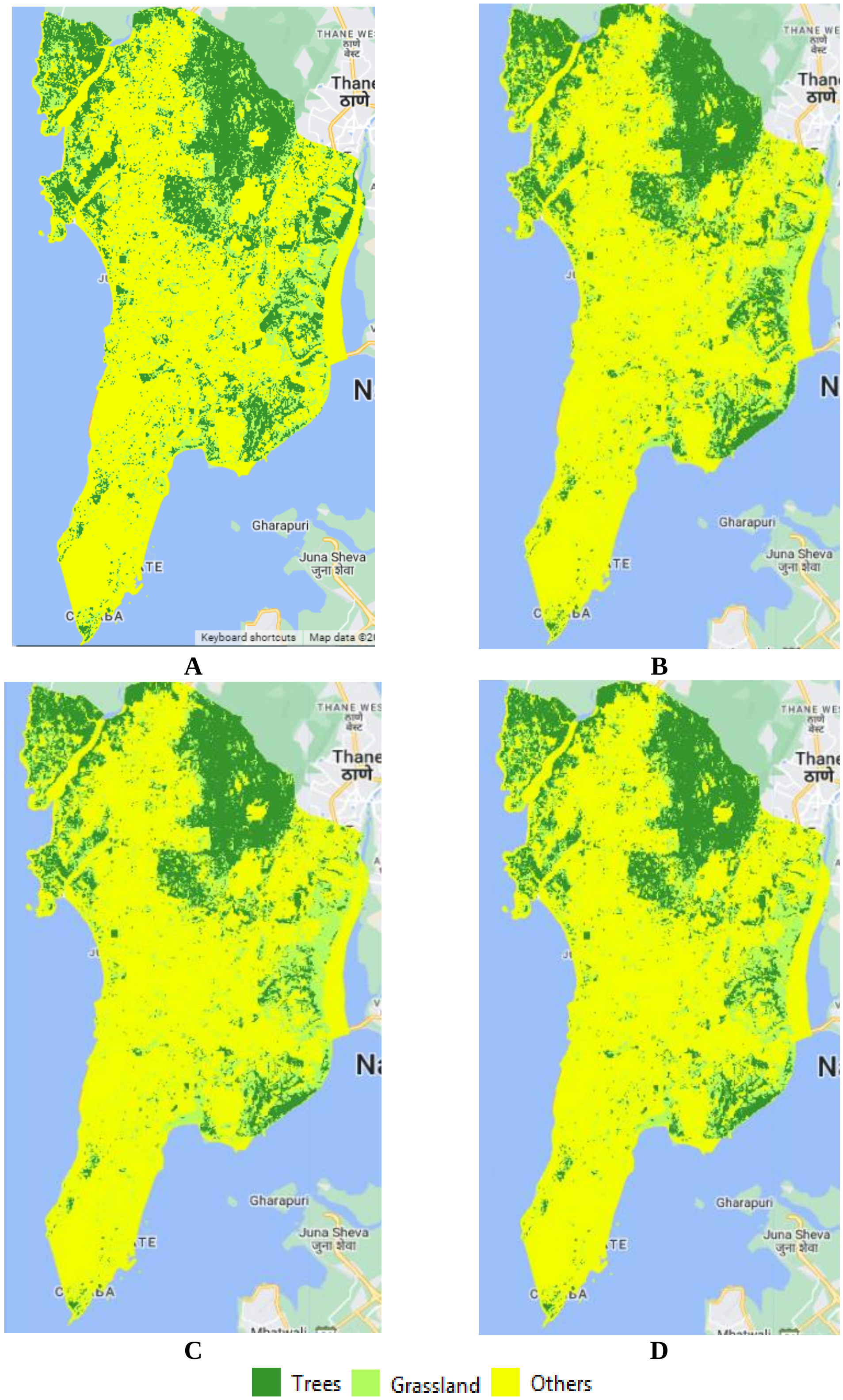

Figure 8. Classification results using (A) original time series, (B) Local-Local approach, (C) Interval-Local approach, and (D) Interval-Interval approach.

5. CONCLUSION & FUTURE DIRECTIONS

This paper proposed a series of time series reconstruction methodologies utilizing temporal interpolations, filtering, and weighted regression concepts, by introducing two different parameters: the nature of temporal window shifts, and the temporal patterns of the reconstructed values. All three methods demonstrated the ability to efficiently restore and reconstruct incomplete time series patterns to different extents, improving the accuracies of the time series data used to carry out classifications on natural land covers over periods of time, as shown in the case study of Mumbai City introduced in the paper. The concept of weighted averages to counteract the atmospheric dampening effects on NDVI and other vegetation signals are also investigated with these methods. The I-L and I-I methods' of incorporating weighted averages of multiple regression lines from interval temporal windows serve as an effective value filter against local abnormal values. Interpolation results from surrounding valid pixels, and an algorithmic weight towards peak values are observed to be effective in preserving both vegetation pixels, and signal variations. The temporal nature of the reconstructed time series is also investigated. While preserving the temporal integrity of the existing time series patterns may be important for certain research where original revisit rates of specific remote sensing products are preferred, in other cases a more consistent temporal interval will produce smoother and more continuous patterns more commonly observed in natural land covers like vegetation.

A

B

C

D

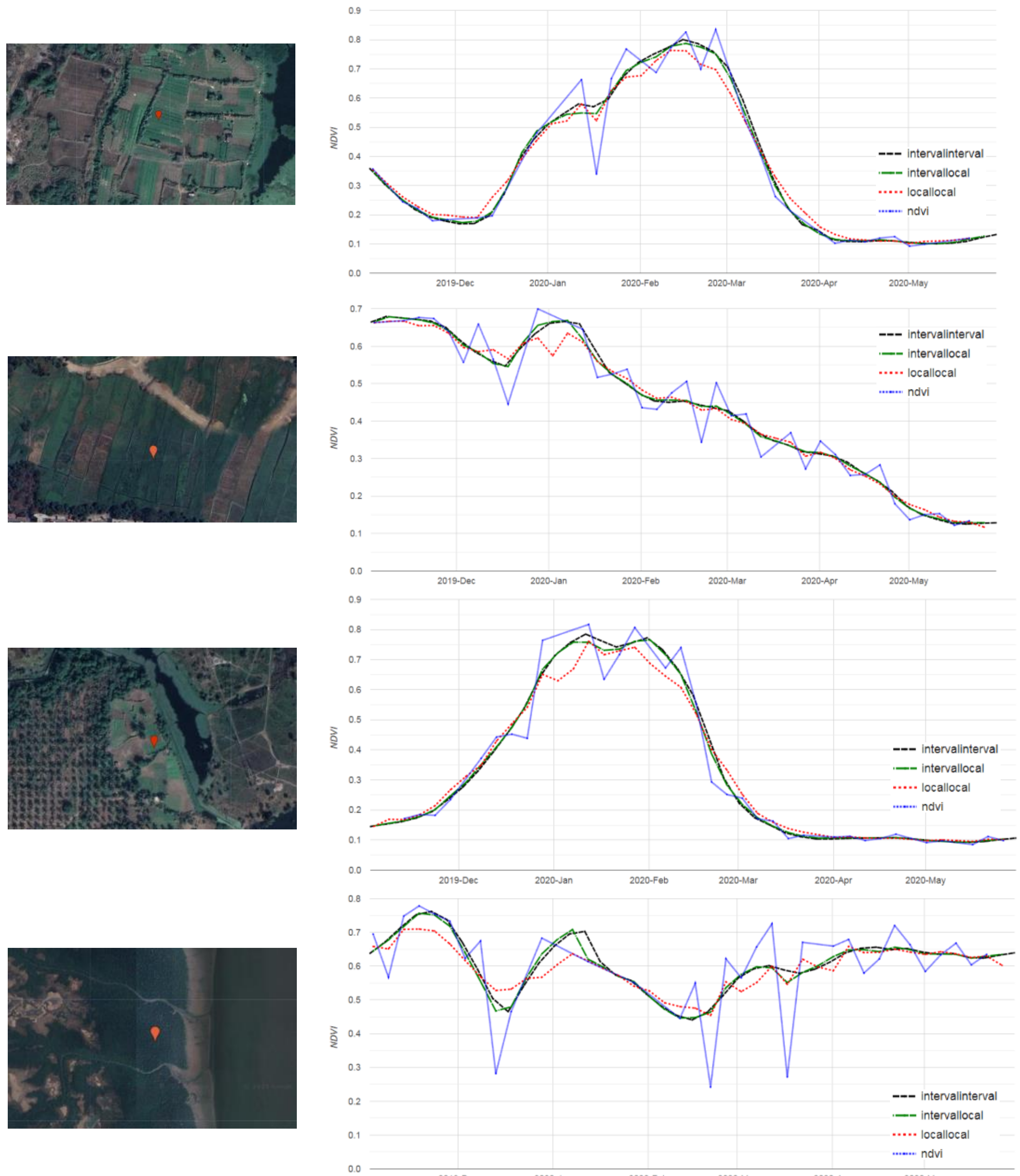

Figure 9. Time series pattern results of different reconstruction methods against original time series.

The current approach is still preliminary, and more can be done to further improve methodologies on different land cover types. The Mumbai City's mangrove forests, for example, see best classification results using the L-L approach instead, due to the shorter timeframe in this case study bolstering the vegetation ratio adopted. Under more extensive research and timeframes considered, the other approaches can be optimized to cater to characteristics of more unique land covers. By drawing focus on specific steps in existing time series reconstruction methodologies, such as the nature of window shifts and resultant temporal intervals, this paper hopes to serve as a stepping stone for future scholarly work and practical applications, to better improve time series reconstruction of remote sensing products in the advent of accessible big data platforms.

ACKNOWLEDGEMENTS

The authors would like to thank the Google Earth Engine team for making this research work possible by providing the necessary tools, algorithms, and datasets on the web browser platform.

REFERENCES

Cao, R., Chen, Y., Shen, M., Chen, J., Zhou, J., Wang, C., Yang, W., 2018. A simple method to improve the quality of NDVI time-series data by integrating spatiotemporal information with the Savitzky-Golay filter. *Remote Sensing of Environment, 217*, pp. 244-257.

Chen, J., Jönsson, P., Tamura, M., Gu. Z, Matsushita, B., Eklundh, L., 2004. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky-Golay filter. *Remote Sensing of Environment, 91*(3-4), pp. 332-344.

Das, M., Ghosh, S. K., 2017. A deep-learning-based forecasting ensemble to predict missing data for remote sensing analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10*(12), pp. 5228-5236.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, *202*, pp.18-27.

Goward, S. N., Markham, B., Dye, D. G., Dulaney, W., Yang, J., 1991. Normalized difference vegetation index measurements from the Advanced Very High Resolution Radiometer. *Remote Sensing of Environment, 35*(2-3), pp. 257-277.

Guillemot, C., Le Meur, O., 2014. Image inpainting: overview and recent advances. *IEEE Signal Processing Magazine, 31*(1), pp. 127-144.

Jakubauskas, M.E., Legates, D.R., Kastens, J.H., 2001. Harmonic analysis of time-series AVHRR NDVI data. *Photogrammetric engineering and remote sensing*, *67*(4), pp.461-470.

Julien, Y., Sobrino, J. A., 2010. Comparison of cloud-reconstruction methods for time series of composite NDVI data. *Remote Sensing of Environment, 114*(3), pp. 618-625.

Shen, H., Li, X., Cheng, Q., Zeng, C., Yang, G., Li, H., Zhang L., 2015. Missing information reconstruction of remote sensing data: a technical review. *IEEE Geoscience and Remote Sensing Magazine*, *3*(3), pp. 61-85.

Swets, D., Reed, B. C., Rowland, J., Marko, S. E., 1999. A weighted least-square approach to temporal NDVI smoothing. In *From image to information: 1999 ASPRS Annual Conference*.

Wang, L., Qu, J. J., Xiong, X., Hao, X., Xie, Y., Che, N., 2006. A new method for retrieving band 6 of Aqua MODIS. *IEEE Geoscience and Remote Sensing Letters*, *3*(2), pp. 267-270.

Xu, L., Li, B., Yuan, Y., Gao, X., Zhang, T., 2015. A temporal-spatial iteration method to reconstruct NDVI time series datasets. *Remote Sensing*, *7*(7), pp. 8906-8924.

Zhang, C., Li, W., Travis, D., 2007. Gaps-fill of SLC-off Landsat ETM+ satellite image using a geostatistical approach. *International Journal of Remote Sensing*, *28*(22), pp. 5103-5122.

Zhou, J., Jia, L., Menenti, M., 2015. Reconstruction of global MODIS NDVI time series: performance of Harmonic ANalysis of Time Series (HANTS). *Remote Sensing of Environment, 163*, pp. 217-228.

Zhou, J., Jia, L., Menenti, M., Gorte, B., 2016. On the performance of remote sensing time series reconstruction methods - A spatial comparison. *Remote Sensing of Environment, 187*, pp. 367-384.