GAP-FILLING MISSING PIXELS OF SENTINEL-2 USING INTERPOLATION AND UNMIXING TO OBTAIN HIGH-QUALITY TIME SERIES DATA

Neha Kailash Nawandar[1], Shaunak Sen[2], and Sivaramakrishnan Janardhanan[3]
Post-doctoral fellow[1], Professor[2,3], Department of Electrical Engineering, IIT Delhi, New Delhi 110016, India
Email: nnawandar@ee.iitd.ac.in[1], shaunak.sen@ee.iitd.ac.in[2], janas@ee.iitd.ac.in[3]

KEY WORDS: Optical remote sensing, gap-filling, Google Earth Engine.

ABSTRACT: Satellite remote sensing is a globally accepted tool for monitoring different ecosystems at varied scales. Microwave, thermal, and optical imagery can be used to obtain geophysical parameters utilized in numerous applications like soil moisture estimation, observing crop phenology, crop identification, mapping soil characteristics, etc. Out of these three types of remote sensing imagery, optical imagery is the core of many satellites. It is freely available from Landsat 8, Sentinel-2, MODIS, etc., to name a few, and at a cost or on request from commercial satellites. Since there is always a trade-off with spatial-temporal-spectral resolution, having high-resolution data in all three categories is impossible. Among others, Sentinel-2 provides high-resolution data every ten days, but optical imagery always suffers due to clouds, and atmospheric pollution. The temporal frequency worsens in such conditions, so addressing this issue is essential. In this paper, we present our work where we target this issue and attempt to fill the affected pixels using interpolation and unmixing to ensure that the previously unusable data is now fit for use. The gaps in the imagery for a region of interest are filled using the data available on other days. Firstly, we set a threshold for the cloud presence percentage, mask the clouds, create a series using the cloud-masked data, and an equivalent empty series. Secondly, for each image in the series, we search for image occurrence before and after the target date within a specified time window and join those images with the reference image. Finally, we apply interpolation, fill the gaps, apply unmixing, and smoothen the images. As a considerable amount of data is involved in the process, the implementation is done on Google Earth Engine (GEE). We have created time series for different ecosystems using the unaffected and gap-filled images. Diverse regions are chosen to highlight the broad applicability of such implementation. Results are obtained for alpine tundra, desert, urban-rural, and tropical rainforest locations, i.e., Joshimath, Thar, New Delhi, and Andaman regions of India. Sample results of original and gap-filled Red and NIR bands of Sentinel-2 for these four regions are presented in the paper.

1. INTRODUCTION

Optical remote sensing utilizes sensors operating in the EM spectrum's visible, near-infrared, and shortwave infrared regions. These sensors collect data by detecting the radiation reflected from the Earth's surface and storing it in images. Landsat 8 OLI/TIRS (Irons 2012), MODIS (Justice 2002), Sentinel-2 (Drusch 2012), etc., provide free optical images at different spatial, temporal, and spectral resolutions. However, the optical data is corrupted by clouds or atmospheric pollution (Makarau 2014). This leads to capturing partial or incorrect data, and missing information worsens the temporal frequency of the satellite imagery. For instance, consider the case of Landsat 8 satellite data when the cloud coverage is high, affecting the pixels, leading to loss of information or a completely unusable image. In such a case, losing data even for a single day will create a gap of ~1 month between two consecutive images. A similar applies to Sentinel-2A satellite data, where missing information worsens the original temporal resolution of 10 days to ~20 days. A crop phenology application, where the farmers/researchers rely on satellite data, might miss crucial information for different growth stages. So, providing a solution to curb losing satellite data is required. To conduct our study, we have utilized Sentinel-2A-based optical imagery. If we look at MODIS, it has a very high revisit frequency and provides optical imagery almost daily. Losing information for a day or two does not have adverse effects. However, this high temporal resolution is always associated with poor spatial resolution, the highest being 250m (Justice 2002), much less than Landsat 8 and Sentinel-2. Such a low-resolution image is unsuitable for applications requiring intricate information. There are solutions to improve the resolution which is beyond the scope of this paper.

Algorithms and techniques for gap-filling satellite images have been around for decades, and still, researchers are coming up with new methodologies to tackle the missing data. Alternative similar pixel approaches utilize spatial, temporal, spectral, and hybrid data to fill the gaps (Mo 2023) in satellite images. Spatial methods use simple interpolation and perform multi-scale segmentation. They use similar information from cloudy and non-cloudy pixels in neighbouring regions. The main assumption is that the gap and cloud-free regions are in close vicinity and similar in terms of statistical and geometrical structures. Kriging falls under the spatial technique, where spatial covariance and its weights are used to interpolate the missing points. Their execution is simple, but the accuracy is low. Such methods are also not capable of filling gaps on a pixel scale. There are geostatistical and regression methods, but the most popular method is interpolation.

The hybrid methods utilize a combination of spatial-temporal-spectral. The spatio-temporal methods (Mo 2023) use spatial characteristics of contextual information and temporal relevance over a time series. An example is window regression. The spectral-temporal techniques are mainly used for multispectral images over a time series. The spatial-spectral-temporal integrates spatial coherence, multispectral features, and time series characteristics and is primarily learning-based.

The spectral-based methods fill the missing values in the incomplete spectral bands using the information from the complete spectral bands. The incomplete spectral bands, in this case, could be due to some fault in the sensors, so very few methods are available. The main limitation of this technique is that it cannot reconstruct gaps at the same location in every spectral band. Unlike the other three, temporal methods utilize information from one or more images to fill the gaps. Its basis is that the images from the same geographic locations with close acquisition dates have strong relevance. Fourier can be used, giving high-quality time series of remote sensing data and making the image usable. Other temporal interpolation methods are spline interpolation (Srivastava 2019), which utilizes a piecewise polynomial, logistic interpolation, and summation of sinusoids. These are used to mask the clouds, followed by filling the gaps, and the main advantage is that the filled data has a natural look. Apart from the above-mentioned, satellite blending (Foster 2011) is also utilized for filling the missing pixels. MODIS and Landsat images are mostly used as their respective sensors share similarities in spectral information, consistency, and complementary specifications.

Changes in the Earth's surface alter the reflectance captured by the satellite. So, the technique must solve the spectral variances that mostly occur due to the observation conditions, variations in the geographical features, and any natural calamity like flood, fire, landslide, etc. Out of these, both other than the natural calamity can be addressed. This paper uses a temporal-based method to fill the Sentinel-2A satellite imagery gaps. It possesses the qualities that any gap-filling method should have. These include being adaptive to gaps of any size, robust to environmental conditions like terrain, climate, land types, etc., and reacting to the disturbances in the time series both accurately and sensitively, easy to perform, fast in computation, and adaptive to global implementation and for different ecosystems. The remaining paper is organized as: Section 2 discusses the methodology, and the algorithm and corresponding results are mentioned in Section 3. The paper is concluded in Section 4 and finally references are mentioned at the end.

2. PROPOSED METHODOLODY

This section discusses the test sites, data sources, and the workflow for which the study is undertaken.

2.1 Test sites and data: Figure 1 shows the locations across India where the algorithm is implemented to obtain high-quality time series data. For our study, we have chosen four climatically distinct locations, i.e., desert, urban-rural, alpine tundra, and tropical rainforest ecosystems. The regions that belong to these ecosystems are Thar, New Delhi, Joshimath, and Andaman and Nicobar. Sentinel-2 data for three years for the mentioned locations is utilized in the study. Related details are given in Table 1.

Table 1. Sites for which the Sentinel-2 Images are used in the study.

Site name	Site type	Site area (km^2)	Bands and Spatial resolution (m)
New Delhi	Urban and rural	1,483	04 and 10
Joshimath	Alpine tundra	2,458	
Thar	Desert	23,000	
Andaman and Nicobar	Tropical rainforest	8,249	

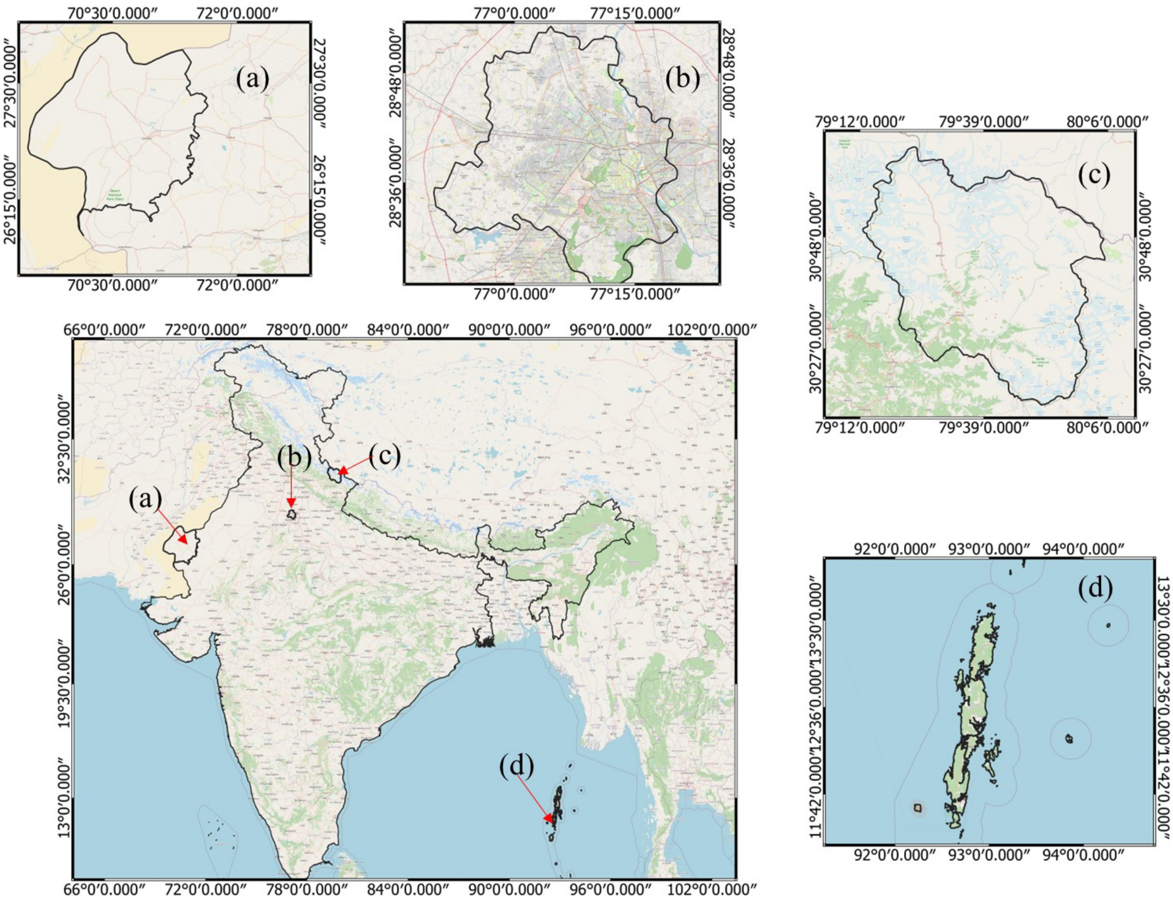


Figure 1. Four different test sites in India

2.2 Algorithm workflow: Figure 2 shows the flowchart for the gap-filling technique presented in the paper. The workflow is primarily divided into three steps. These are (i) selection of data, (ii) filtering the data, and (iii) gap-filling using interpolation, unmixing, and smoothening. As mentioned earlier, Sentinel-2 satellite imagery is used in the study due to its higher spatial resolution in Visible and near-IR bands and temporal frequency compared to Landsat 8 and MODIS imagery. The entire dataset is initially loaded and filtered per the region of interest and required timeline. It is evident that the optical images are rarely free of clouds, and their pixels are contaminated depending on the cloud cover at the time of capture. An additional cloud cover and shadow percentage threshold are set as a precautional measure, and the pixels above this value are automatically masked. The same is followed for snow cover, which is specific to snow-clad regions, and in our study, it applies to the Joshimath region. Thus, in all, the algorithm not only fills the gaps but also estimates such pixel values. It is done to avoid any erroneous values during interpolation.

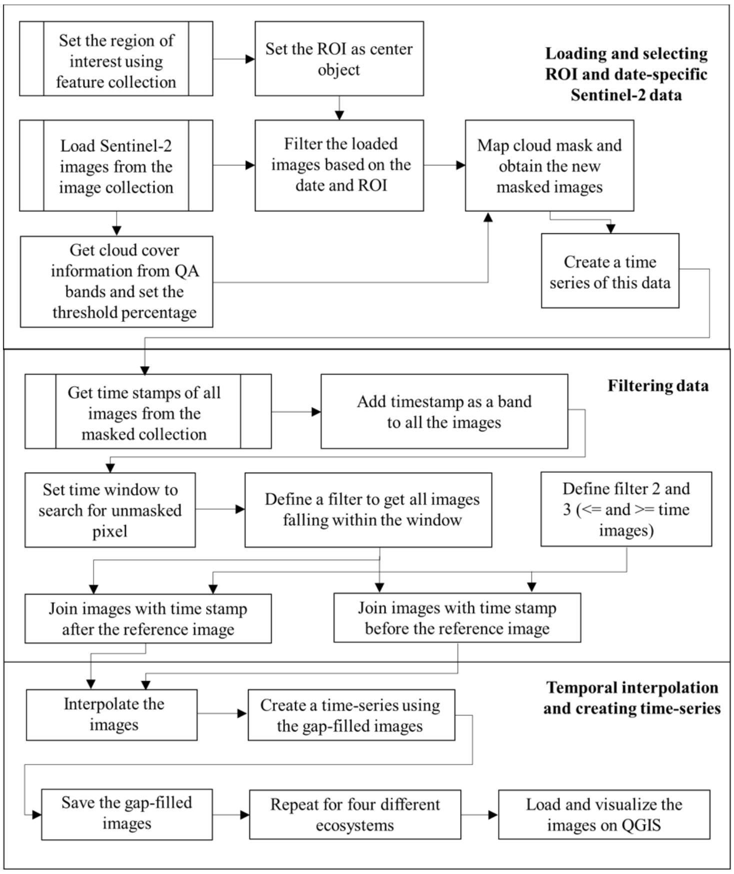


Figure 2. Flowchart of the gap-filling technique used for generating high-quality time series data.

Algorithm 1: Gap-filling pseudo code

1: **Require**: S2A → Sentinel-2A dataset
2: **Create**: t.shp, d.shp, a.shp, j.shp → shape files for four ecosystems
3: **Set**: c_t, sh_t, s_t → cloud cover, shadow, and snow thresholds
4: **Output**: $S2A_f$ → gap filled S2A
5: **for** im **in** S2A:
   restrict im to ROI, time range and get metadata for im
   p in im ∀ p∈{c_t, sh_t, s_t} → search for pixels with values above thresholds in step 3
   mask p
   find im_t for unmixing → image at 't' without cloud/shadow/snow
   get {t}∀ im and attach to im → get timestamps and attach to each image
   ρ ∀ im → surface reflectance from DN
   unmix ∀ im
6: Obtain $S2A_{t_s}$:{im_{t_1}, im_{t_2} ... im_{t_n}} → time series of images with masked pixels and timestamps
7: Set T → time window to search for valid pixel
8: f_1 and f_2 → filters to use on reference im, i.e., image to be filled
9: Get t_1 and t_2 from T for im from $S2A_{t_s}$ → before and after the reference date
10: Get im_{t_1} and im_{t_2} → data at t_1 and t_2 for interpolation
11: Compute $^{t-t_1}/_{t_2-t_1}$ → time ratio (tr)
12: Compute $y_1+(y_2-y_1)*tr$ → the interpolated image
13: Get the distance → parameter for smoothening the image
14: Apply the filter → smoothened image
15: Get interpolated time series $S2A_f$
16: End

The images are pre-processed after masking the cloud-covered, snow-covered, and shadow pixels. In this pre-processing step, DN images are converted to surface reflectance, and timestamps are obtained from the metadata file accompanying the images. A time series collection comprising the images converted to surface reflectance with the timestamps attached to each image is created. For every missing pixel of each image in the series, corresponding pixels with valid values are searched for. Such a collection is created for every image and is further used during the interpolation step.

Now, after the images are searched, their corresponding timestamps are used to compute the time ratio (tr). This can be seen from the step 10 of Algorithm 1. This is followed by estimating the interpolated image for the original reference image with gaps. Next, distance is computed which is required for smoothening the image. Finally, filter is applied to obtain the gap-filled smoothened image. Similar is repeated for every image in the original time series to obtain a gap filled series. To showcase the wide applicability of the algorithm, results for different ecosystems are presented in Section 3.

3. RESULTS AND DISCUSSION

This section presents the results obtained after applying the methodology to Sentinel-2A data for four different ecosystems. The ecosystems are desert, urban-rural, alpine tundra, and tropical rainforest. These include, Thar, New Delhi, Joshimath, and Andaman and Nicobar regions of India.

Figure 3 shows the original and gap filled Red band images for these four locations. As seen from the Figure 3, sample results for two days in the test time window are shown. Here, each row corresponds to either the original image or the gap-filled image for the test location. The results clearly showcase the difference between the original and the output image. For instance, let us consider the Thar region and the result shown in the above row. Here, it can be observed that there is almost no data available on the reference day, still the output gap-filled image has information at many locations. This is obtained using the information from the images available on other days. Similar can be observed for the other three locations as well.

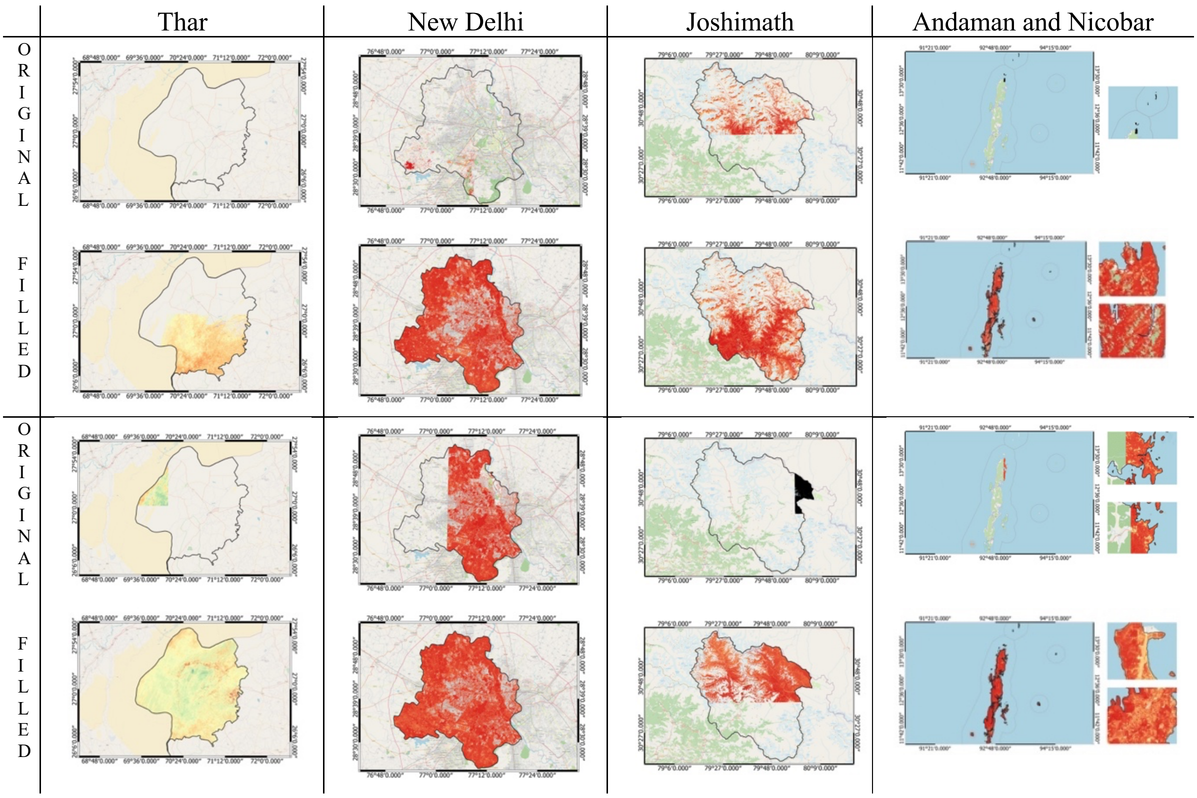


Figure 3. Sample images of the original and gap filled Red band for the different test sites

The methodology is applied, and results are also obtained for the NIR band. The corresponding results are shown in Figure 4. The observations for this case are similar to what we have observed for the Red band. We have presented results for Red and NIR bands only, but the algorithm was applied to all the Sentinel-2A optical bands. An NDVI time series for the target duration is also obtained using both the original and the gap-filled images. Results for Delhi and Joshimath regions are shown in Figure 5.

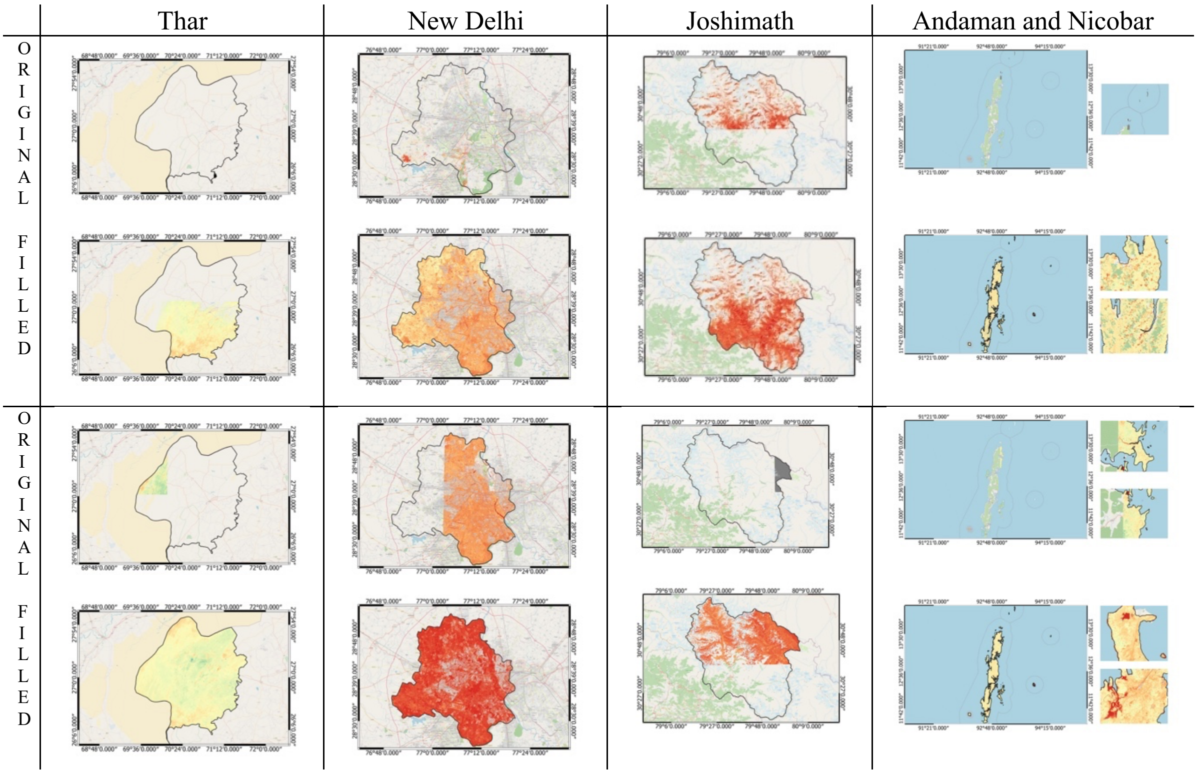
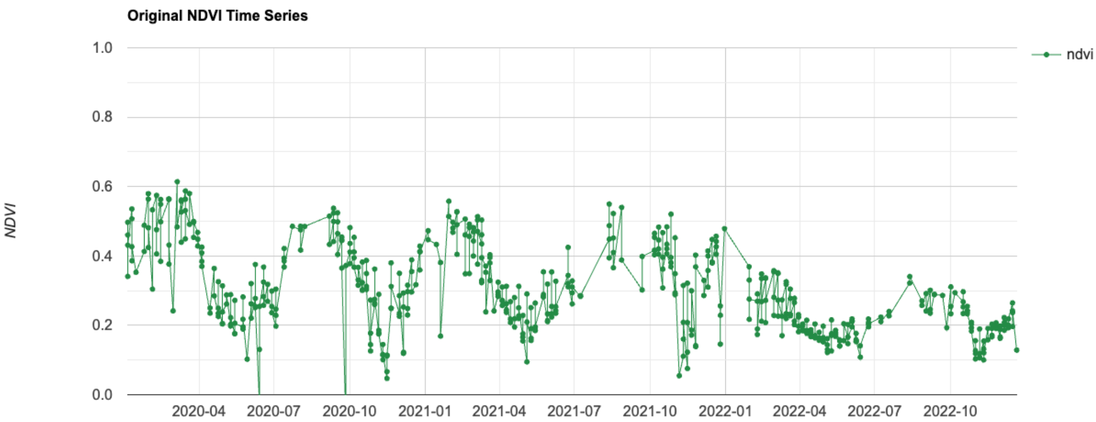
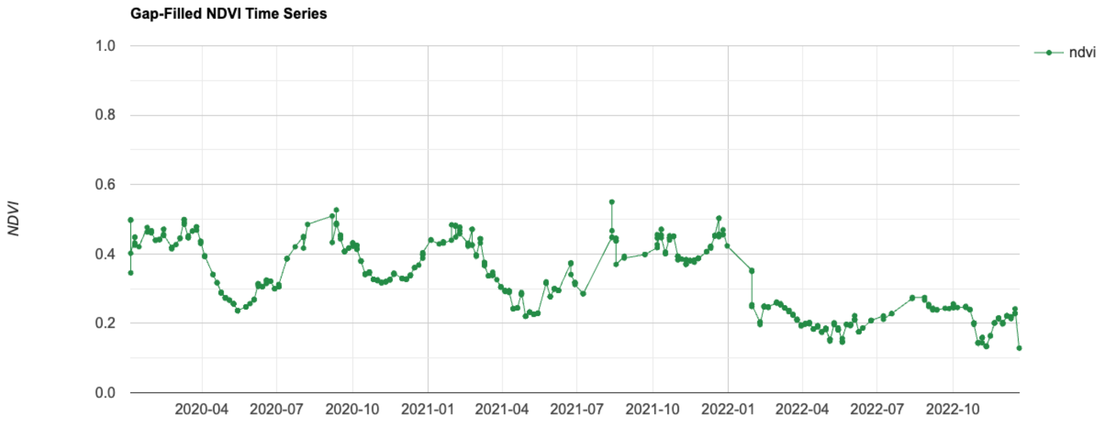


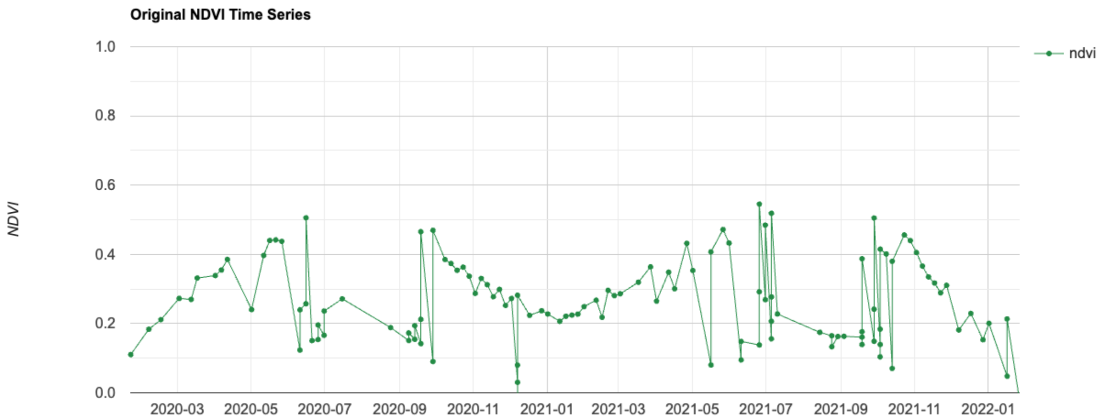
Figure 4. Sample images of the original and gap filled NIR band for the different test sites



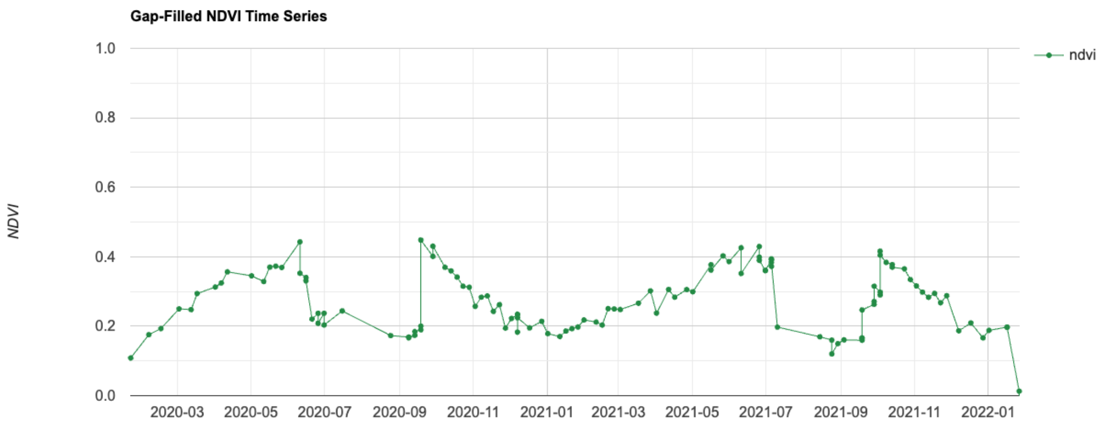
(a) Original NDVI time series for Delhi



(b) Gap filled NDVI time series for Delhi



(c) Original NDVI time series for Joshimath



(d) Gap-filled NDVI time series for Joshimath

Figure 5. Original and gap-filled NDVI time series for two sites.

4. CONCLUSION AND FUTURE SCOPE

4.1 Conclusion: There are gaps in satellite optical imagery, possibly due to cloud contamination, sensor malfunctioning, or atmospheric pollution. This hinders the study and analysis of numerous applications that require complete optical imagery directly or fuse it with other satellite data. Filling such missing pixels of satellite optical images is the first step towards multi-sensor data fusion to obtain high-quality time series data. For the study presented in this paper, the Sentinel-2A satellite was chosen over Landsat 8 OLI/TIRS and MODIS because of its spatial resolution in our targeted wavelength region and temporal frequency. So, this paper discussed a technique to fill the missing and affected pixels of Sentinel-2 images. Gaps were filled using temporal interpolation with the help of data available on other days and unmixing with the help of LULC. A threshold for cloud cover was set for the initial data screening, and the corresponding pixels were eliminated to prevent incorrect computations during interpolation. Then, a data series was created using the modified images. For each image in the series, the algorithm searched for images with non-zero and non-nan pixels for gap-filling. The search was done in a specific time window, and these images were used for interpolating and filling the missing pixels in the reference image. Finally, a time series of the gap-filled images was obtained and smoothened. The paper discussed the results and time series obtained for four ecosystems: desert, urban-rural, alpine tundra, and tropical rainforest. Locations from different ecosystems were chosen to showcase the versatility of such techniques. The complete implementation was done on Google Earth Engine (GEE) due to the huge data involved in the study. Successful implementation and results obtained using such techniques make the previously hazy data fit for usage, allowing researchers to study and analyze decades of satellite data.

4.2 Future scope: The study presented in the paper fills Sentinel-2A satellite imagery gaps using data captured on other days from the same satellite. However, if the missing pixels are detected in consecutive images, it is difficult to correctly reconstruct them. In such a scenario, satellite data fusion could be a better solution. This ensures filling the missing pixels and blending the data to obtain a reliable dataset. Our future work targets creating a gap-free time series of daily surface reflectance products for agricultural and other applications.

Acknowledgments
There are no acknowledgments for this submission.

References:

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P. and Meygret, A., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment*, *120*, pp.25-36.

Foster, J.L., Hall, D.K., Eylander, J.B., Riggs, G.A., Nghiem, S.V., Tedesco, M., Kim, E., Montesano, P.M., Kelly, R.E., Casey, K.A. and Choudhury, B., 2011. A blended global snow product using visible, passive microwave and scatterometer satellite data. *International journal of remote sensing*, *32*(5), pp.1371-1395.

Irons, J.R., Dwyer, J.L. and Barsi, J.A., 2012. The next Landsat satellite: The Landsat data continuity mission. *Remote sensing of environment*, *122*, pp.11-21.

Justice, C.O., Townshend, J.R.G., Vermote, E.F., Masuoka, E., Wolfe, R.E., Saleous, N., Roy, D.P. and Morisette, J.T., 2002. An overview of MODIS Land data processing and product status. *Remote sensing of Environment*, *83*(1-2), pp.3-15.

Makarau, A., Richter, R., Müller, R. and Reinartz, P., 2014. Haze detection and removal in remotely sensed multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(9), pp.5895-5905.

Mo, Y., Xu, Y., Liu, Y., Xin, Y. and Zhu, S., 2023. Comparison of gap-filling methods for producing all-weather daily remotely sensed near-surface air temperature. *Remote Sensing of Environment*, *296*, p.113732.

Srivastava, P.K., Pandey, P.C., Petropoulos, G.P., Kourgialas, N.N., Pandey, V. and Singh, U., 2019. GIS and remote sensing aided information for soil moisture estimation: A comparative study of interpolation techniques. *Resources*, *8*(2), p.70.