

LEAF AREA INDEX TIME SERIES IMPUTATION FOR EARLY YIELD PREDICTION

Christoph Jörges¹, Jens E. d'Hondt², Georgios Chatzigeorgakidis³, Silke Migdall¹, Christian Miesgang¹,
Susanne Karg¹, Heike Bach¹, Panagiotis Betchavas³, Dimitrios Skoutas³

¹VISTA – Remote Sensing in Geosciences GmbH, Gabelsbergerstrasse 51, 80333 Munich, Germany,

²TU/e – Eindhoven University of Technology, De Zaale 1, 5612 AZ Eindhoven, the Netherlands,

³IMSI, Athena Research Center, Artemidos 6 & Epidavrou, 15125 Marousi, Greece

ABSTRACT

Leaf Area Index (LAI) is a key parameter in crop growth models, and its accurate estimation is crucial for yield prediction. However, LAI data values are often missing or incomplete due to various reasons, such as sensor failures or cloud cover. In this paper, we propose a set of time series data imputation methods for LAI values derived from satellite images by radiative transfer model (RTM) inversion. The methods perform temporal interpolation either at the level of individual pixels or on spatial aggregates. Our experimental evaluation demonstrates that our approach can be applied to various crop types and has the potential to improve the accuracy and timeliness of yield prediction.

Index Terms— leaf area index, time series, field segmentation, data imputation, yield prediction

1. INTRODUCTION

Predicting crop growth and yield development is crucial both on the local level for farm management measures, as well as on the regional to continental level to predict food supply and allow early warnings for food shortages due to yield loss. Climate change impacts on the environmental conditions for agriculture further exacerbate the risk of yield loss and therefore food security. Earth Observation (EO) data combined with modeling can provide yield predictions on different spatial scales, continuously and securely using an open and independent source. Public services usually use medium resolution EO data and bring out monthly reports on potential food shortages on national and/or continental scale, mainly targeting other public service providers. Conversely, commercial services like YPSILON^{®1} use high resolution optical imagery and start their predictions 6 to 8 weeks before harvest, targeting commercial customers. However, for farmers, predictions from both parties often come late in season. To effectively influence crop development with smart farming measures, predictions need to be performed early in the season while the

crops are still in the vegetative phase. Our work in this paper is part of our ongoing efforts to address this challenge in the context of the EU Horizon Europe project STELAR².

An important first step in obtaining earlier yield predictions is to improve the quality of the input data. For this, effective data imputation methods need to be developed to close data gaps and provide input for physical modeling and machine learning (ML) techniques that derive crop type, crop status, and current crop growth [9]. In this paper, we focus on the imputation of the leaf area index (LAI), which – as a proxy for crop growth over time – is a key parameter for yield predictions. By radiative transfer model (RTM) inversion, the parameter can be retrieved from the multispectral reflectance signals of satellite data [8]. Nevertheless, daily, cloud-free acquisitions are not available, leading to uncertainties in crop growth modeling if data are missing over long periods. Therefore, the objective of this study is to present our current work on missing values imputation methods for LAI time series derived from Sentinel-2 via RTM inversion. The data cover multiple crop types and seasons. The methods are divided into two approaches: temporal interpolation and temporal interpolation on spatial aggregates like fields. Experiments with LAI data near Bordeaux, France are performed to evaluate both approaches. Future work involves analyzing the impact of improved imputation on early yield predictions.

2. RELATED WORK

Missing values imputation. Time series missing values imputation is a crucial task in many fields, including finance, economics, and environmental science. A variety of methods have been proposed, including matrix-based and pattern-based ones [5]. The former attempt to infer missing time series blocks by representing the original data using matrices and then applying decomposition. Recovering the original time series produces values at each timestamp of each gap. SVDImpute [12] is a popular such method, which selects the k most significant columns of a decomposed matrix and uses linear combination to infer the missing values. Pattern-

ACKNOWLEDGEMENTS: This work has received funding from the European Union's Horizon Europe project STELAR (Grant No. 101070122).

¹<https://epsilon.services/>

²<https://stellar-project.eu/>

based methods assume profound similarities among the time series and attempt to fill in a gap using similar reference series. A popular pattern-based method is DynaMMo [6], which detects and uses co-evolving patterns among the time series in a dataset. Techniques that employ neural networks have also been proposed [2]. However, in this paper, we focus on matrix- and pattern-based techniques since the series in a dataset tend to be linearly dependent on one another.

Field delineation. Agricultural land-use statistics often offer more insightful and stable information when analyzed at the level of individual fields rather than individual pixels [11]. However, obtaining field boundary maps covering large areas with potentially thousands of farms is a challenging task. Field delineation, also known as crop field boundary detection, aims at developing automated techniques that can extract parcel boundaries directly from satellite images. Generally, the approaches to this problem can be partitioned into classical computer vision (CV) methods and ML approaches. Classic CV methods generally rely on raster analysis and employ several image processing techniques such as blurs, filter, and edge detection algorithms to detect fields [1, 10, 11, 14]. ML approaches arose in recent years for addressing the limitations of classic edge-based and region-based methods; their sensitivity to noise and excessive parameterization requiring manual tweaking and leading to context-specific results [4, 15]. Notably, Waldner et al. [15] addressed these issues by mapping the task to a multi-task semantic segmentation problem and utilizing the state-of-the-art ResUNet architecture to identify the extent, boundaries, and edge-distances of fields. This greatly improved model generalization while essentially being parameter-free.

3. DATA CHARACTERISTICS

The study area covers a region of about 10,004 km² south-west of Bordeaux, France, which corresponds to the Sentinel-2 tile 30TYQ. The area is dominated by agricultural land use and forests in the south-western parts. Multispectral space-borne data of Sentinel-2 Level 1C from 2020 to 2022 are used as input. After specific cirrus and atmospheric corrections designed for agricultural parameter retrieval in Europe, cloud and shadow masking was performed. Also, a preliminary land cover classification for snow, water, and vegetation, and masking of forests and urban areas was conducted. The LAI values were received by crop type independent derivation via RTM inversion with the Soil-Leaf-Canopy (SLC) model [13].

The above procedure outputs a set of 231 raster images (10002 × 10002 pixels) with LAI values for every valid vegetation pixel in the period from Jan. 2020 to Dec. 2022. Time series of LAI values were extracted from this array by slicing over the time dimension, masking missing and faulty LAI values with NaN tokens. Fig. 1a depicts an example time series, containing three gaps, with a total of 53 missing values. Figure 1b shows the distribution of missing values for both

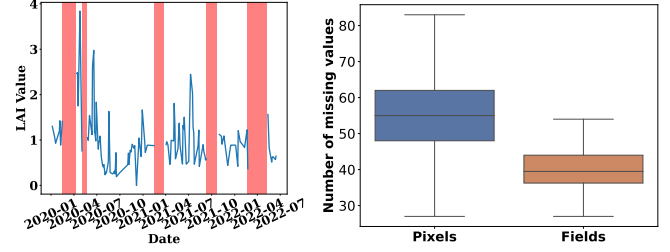


Fig. 1: (a) Random LAI time series with exemplary missing values.; (b) Boxplot of missing value statistics.

pixel-level and field-level time series.

4. IMPUTATION METHODS

4.1. Pixel-level Temporal Interpolation

We present a novel approach for temporal missing values imputation named cMVI (i.e., combined Missing Values Imputation) that attempts to combine the outputs of 12 state-of-the-art methods evaluated in [5] to yield better imputation performance. These include SVDImpute, SoftImpute, SVT, CDRec, GROUSE, SPIRIT, ROSL, TRMF, TeNMF, TKCM, DynaMMo, and STMVL. They take as input a set of time series and attempt to fill in gaps in each one based on both temporal information and other time series values in the set.

To optimize information sharing between time series, we initially apply k -means clustering, to obtain N clusters of similar time series (we found 5 to be optimal in our case). Then, for each cluster, we split the set of time series (in the time domain) into a training and a validation set and introduce custom gaps in the training set. Consequently, we apply the 12 imputation methods on the training set and use an XGBoost regressor to compose a final prediction based on the obtained estimations, using the actual LAI value as a ground truth. This way, we effectively create an ensemble of imputation methods.

4.2. Field-level Temporal Interpolation

As discussed in Section 2, processing regional crop and land use statistics on a *field-level* makes sense from both an efficiency and quality perspective. First, clustering pixels can heavily reduce the computational cost of all downstream tasks (e.g., imputation) as it decreases the number of time series to be processed. Second, the effectiveness of classification and prediction methods can be enhanced when fields are processed together as a single entity, since they can be characterized more robustly by its average reflectance and additional characteristics like size, shape, and texture [11].

As such, a second temporal interpolation approach was developed, including field delineation as a pre-processing step to extract clusters of pixels corresponding to crop fields. Then,

the methods of Section 4.1 are performed on the aggregated LAI values of the cluster over time. We opted for using the median values, but note that other aggregation methods could also work in certain contexts. For field delineation, the ReSUNet model architecture of Waldner et al. [15] was used as it provided robust results and proved to generalize over multiple countries in the EU. The segmentation model was trained and validated on the AI4Boundaries dataset [3], which is an open AI-ready dataset for field delineation on Sentinel-2 data, including ground truth vector shapes.

Performing inference with our model on the study area resulted in a dataset of 133,382 fields with corresponding (aggregated) time series. This implies a 750x reduction of the total search space, compared to pixel-level imputation. Figure 2c shows the output of the model for an image patch.

5. EVALUATION

In this section, we present our preliminary results of temporal missing values imputation on individual pixels and on spatially aggregated data.

5.1. Experimental Setup

The following imputation approaches are compared on sets of time series extracted from the data described in Section 3.

Baseline. Simple linear interpolation (SLI).

Pattern-adjusted SLI. We perform weighted linear interpolation between supporting data points. On this, reference time series of different crop types (spring barley, winter rapeseed, and winter wheat) of the years 2020-2022 are used for weighted adjustment of the linear interpolation. The reference time series come from numerical model assimilation with the crop growth model PROMET [7] of selected aggregated 4×4 km pixels in the study area. The supporting data points are not modified within this approach, while a moving average smoothing is performed afterwards.

cMVI-Plain. We apply our cMVI algorithm on the whole length of plain LAI time series, without any pre-processing.

cMVI-Smooth. We apply our cMVI algorithm on the whole length of LAI time series, after applying Savitzky–Golay smoothing on each one.

cMVI-PastYear. Due to climate change, LAI timeseries generally have decreased autocorrelations, making recent values less related to the values from previous years. To investigate this impact, we apply our cMVI algorithm only on the last year of LAI time series, after applying Savitzky–Golay smoothing on each one.

The approaches were evaluated on two datasets; one including the LAI values over time for 1000 individual pixels (i.e., pixel-level), and one including the aggregated LAI values of 1000 fields (i.e., field-level). The field-level data were constructed by extracting the top-1000 fields with the least

Method	Pixel-level		Field-level	
	MAE	RMSE	MAE	RMSE
cMVI-Plain	0.69	0.95	0.69	0.92
cMVI-Smooth	0.47	0.68	0.45	0.68
cMVI-1Year	0.39	0.61	0.37	0.59
Pattern-adjusted SLI	0.48	0.71	N/A	N/A
Baseline	0.49	0.74	0.51	0.76

Table 1: Accuracy metrics for both imputation methods

amount of missing values. The pixel-level data were constructed by extracting the pixel with the least amount of missing values for each field in the field-level data. This was to ensure a fair comparison between pixel-level temporal imputation and field-level temporal imputation.

5.2. Interpolation results

Fig. 2a depicts the imputed time series (the gap is illustrated using a red-shaded rectangle) of an example pixel using the different temporal imputation approaches, along with the ground truth. Notice that, all approaches correctly estimate the LAI value decrease. Compared to the baseline, the rest of the methods manage to follow the overall trend of the original time series, with cMVI-1Year outperforming the best. The cMVI-Smooth approach also outperforms cMVI-Plain, indicating that smoothing the data can improve accuracy, as possible outliers are ignored.

Fig. 2b shows an imputed time series containing the median LAI values of an evaluation field. In this case, the cMVI-Smooth approach manages to closely match the original series. Interestingly, cMVI-1Year overestimates the LAI value rise after 2022-05-26, possibly due to the fewer data available for imputation. Note that the pattern-adjusted SLI method was not used for imputation at the field level, since it was not feasible to find patterns for aggregated synthetic pixels.

Table 1 contains the overall results of our study; specifically, the average mean absolute error (MAE) and root mean squared error (RMSE) over all pixels and fields for pixel-level and field-level imputation, respectively. In the case of pixel-level imputation, the cMVI-1Year approach outperformed the competition, indicating that the data from the previous year may be more suitable for imputation, since more similarities are expected in the crop type, weather conditions, etc. Note that, the cMVI-Plain approach performed worse than both the baseline and pattern-adjusted SLI; this confirms the advantage of performing a priori smoothing on the data. On the field-level, the results are similar but slightly improved in almost all cases. This might be attributed to the dampened noise resulting from aggregating the individual pixel values, which increases the accuracy of the imputation methods.

6. CONCLUSIONS AND FUTURE WORK

This paper presents our work on pixel-level and field-level missing values imputation for time series of LAI data. The

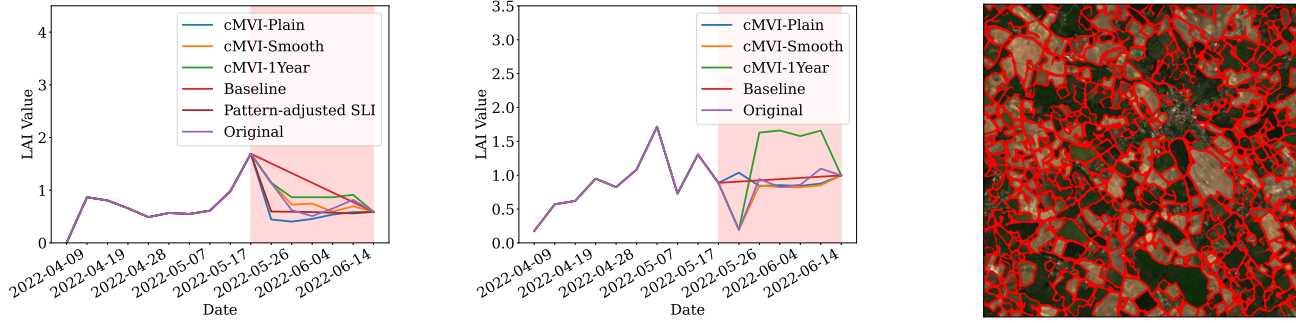


Fig. 2: (a) Pixel-level imputation (b) Field-level imputation (c) Field delineation through image segmentation.

proposed approaches showed superior performance compared to simple imputation methods, potentially impacting the accuracy and timeliness of yield prediction, which is critical for effective crop management and food security.

The presented results indicate that ensembles of imputation techniques can improve performance. By aggregating time series the imputation performance was further improved. While these techniques extract patterns from the (training) data to improve imputation, they are inherently *temporal*, meaning they only consider temporal patterns in the time series they ought to impute. Spatio-temporal imputation, on the other hand, extends this concept by also leveraging spatial patterns in the data by including other, neighbouring time series. Due to the usage of field delineation leading to a stark search space reduction, these methods become feasible and should be considered in future work. Furthermore, future work can also focus on exploring the integration of other data sources, such as weather, soil data, and hyperspectral satellite data, to improve imputation performance of LAI values. This task is often referred to as *multiple imputation* or *data fusion*.

REFERENCES

- [1] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.
- [2] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *NeurIPS 18*, 2018.
- [3] R. d’Andrimont, M. Claverie, P. Kempeneers, D. Muraro, M. Yordanov, D. Peressutti, M. Batič, and F. Waldner. Ai4boundaries: an open ai-ready dataset to map field boundaries with sentinel-2 and aerial photography. *Earth System Science Data*, 15(1):317–329, 2023.
- [4] A. García-Pedrero, C. Gonzalo-Martín, and M. Lillo-Saavedra. A machine learning approach for agricultural parcel delineation through agglomerative segmentation. *International Journal of Remote Sensing*, 38(7):1809–1819, 2017.
- [5] Mourad Khayati, Alberto Lerner, Zakhar Tymchenko, and Philippe Cudré-Mauroux. Mind the gap: an experimental evaluation of imputation of missing values techniques in time series. In *VLDB 20*, pages 768–782, 2020.
- [6] Lei Li, James McCann, Nancy S Pollard, and Christos Faloutsos. Dynammo: Mining and summarization of coevolving sequences with missing values. In *SIGKDD 09*, pages 507–516, 2009.
- [7] Wolfram Mauser and Heike Bach. Promet - large scale distributed hydrological modelling to study the impact of climate change on the water flows of mountain watersheds. *Journal of Hydrology*, 376(3-4):362–377, 2009.
- [8] Silke Migdall, Heike Bach, Jans Bobert, Marc Wehrhan, and Wolfram Mauser. Inversion of a canopy reflectance model using hyperspectral imagery for monitoring wheat growth and estimating yield. *Precision Agriculture*, 10(6):508–524, Dec 2009.
- [9] Silke Migdall, Sandra Dotzler, Eva Gleisberg, Florian Appel, Markus Muerth, Heike Bach, Giulio Weikmann, Claudia Paris, Daniele Marinelli, and Lorenzo Bruzzone. Crop water availability mapping in the danube basin based on deep learning, hydrological and crop growth modelling. *Engineering Proceedings*, 9(1), 2021.
- [10] Ramakant Nevatia and K Ramesh Babu. Linear feature extraction and description. *Computer Graphics and Image Processing*, 13(3):257–269, 1980. ISSN 0146-664X.
- [11] Heather C. North, David Pairman, and Stella E. Belliss. Boundary delineation of agricultural fields in multitemporal satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(1):237–251, 2019.
- [12] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.
- [13] Wout Verhoef and Heike Bach. Coupled soil–leaf–canopy and atmosphere radiative transfer modeling to simulate hyperspectral multi-angular surface reflectance and toa radiance data. *Remote Sensing of Environment*, 109(2):166–182, 2007.
- [14] Matthias Wagner and Natascha Oppelt. Extracting agricultural fields from remote sensing imagery using graph-based growing contours. *Remote Sensing*, 12:1205, 04 2020.
- [15] François Waldner and Foivos I. Diakogiannis. Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network. *Remote Sensing of Environment*, 245:111741, 2020. ISSN 0034-4257.