Reconstruction of Sentinel-2 Derived Time Series Using Robust Gaussian Mixture Models - Application to the Detection of Anomalous Crop Development

Florian Mouret, Mohanad Albughdadi, Sylvie Duthoit, Denis Kouamé, Guillaume Rieu, Jean-Yves Tourneret

Reconstruction of Sentinel-2 Derived Time Series Using Robust Gaussian Mixture Models —Application to the Detection of Anomalous Crop Development⋆

Florian Mouret[a,b,*], Mohanad Albughdadi[a], Sylvie Duthoit[a], Denis Kouamé[c], Guillaume Rieu[a] and Jean-Yves Tourneret[b]

[a] *TerraNIS, 12 Avenue de l'Europe, 31520 Ramonville-Saint-Agne, France*
[b] *University of Toulouse / IRIT-INP-ENSEEIHT / TéSA, 2 Rue Charles Camichel, 31000 Toulouse, France*
[c] *University of Toulouse / IRIT-UPS, 118 Route de Narbonne, 31062 Toulouse Cedex 9, France*

ARTICLE INFO

Keywords:
Crop monitoring
Sentinel-1
Sentinel-2
Multispectral
Synthetic Aperture Radar (SAR)
Isolation Forest
Anomaly detection
Heterogeneity
Vigor
Missing data
Expectation-Maximization
Robust Gaussian Mixture Model
Data imputation

ABSTRACT

Missing data is a recurrent problem in remote sensing, mainly due to cloud coverage for multispectral images and acquisition problems. This can be a critical issue for crop monitoring, especially for applications relying on machine learning techniques, which generally assume that the feature matrix does not have missing values. This paper proposes a Gaussian Mixture Model (GMM) for the reconstruction of parcel-level features extracted from multispectral images. A robust version of the GMM is also investigated, since datasets can be contaminated by inaccurate samples or features (e.g., wrong crop type reported, inaccurate boundaries, undetected clouds, etc). Additional features extracted from Synthetic Aperture Radar (SAR) images using Sentinel-1 data are also used to provide complementary information and improve the imputations. The robust GMM investigated in this work assigns reduced weights to the outliers during the estimation of the GMM parameters, which improves the final reconstruction. These weights are computed at each step of an Expectation-Maximization (EM) algorithm by using outlier scores provided by the isolation forest (IF) algorithm. Experimental validation is conducted on rapeseed and wheat parcels located in the Beauce region (France). Overall, we show that the GMM imputation method outperforms other reconstruction strategies. A mean absolute error (MAE) of 0.013 (resp. 0.019) is obtained for the imputation of the median Normalized Difference Index (NDVI) of the rapeseed (resp. wheat) parcels. Other indicators (e.g., Normalized Difference Water Index) and statistics (for instance the interquartile range, which captures heterogeneity among the parcel indicator) are reconstructed at the same time with good accuracy. In a dataset contaminated by irrelevant samples, using the robust GMM is recommended since the standard GMM imputation can lead to inaccurate imputed values. An application to the monitoring of anomalous crop development in the presence of missing data is finally considered. In this application, using the proposed method leads to the best detection results, especially when SAR data are used jointly with multispectral images. Exploiting the information contained in cloudy multispectral images instead of removing these images is beneficial for this application.

1. Introduction

Remote sensing images have become an essential tool for many agricultural applications, including precision farming Weiss et al. (2020), primarily because they can be used to provide valuable information about vegetation without a need for on-site visits (Schulz et al., 2021). In recent years, the amount of freely accessible remote sensed images has drastically increased, especially thanks to the Copernicus mission operated by the European Space Agency

⋆This document is the results of the research project funded by TerraNIS SAS. and ANRT (convention CIFRE no. 2018/1349). It was partly supported Minciencias through the 910-2019-Program ECOS Nord entitled "Fusion of synthetic aperture radar and optical multitemporal imaging and its application to anomaly detection in citrus" under Grant 8598.

⋆⋆This is a preprint version of the research paper published in Elsevier Computers and Electronics in Agriculture: https://doi.org/10.1016/j.compag.2022.106983.

*Corresponding author
florian.mouret@terranis.fr / florian.mouret@irit.fr (F. Mouret)
ORCID(s): 0000-0003-3719-6767 (F. Mouret)

(ESA). Its first multispectral high resolution satellite (Sentinel-2A) was launched in 2015, followed by a second satellite in 2017 (Sentinel-2B) (Drusch et al., 2012). Two synthetic aperture radar (SAR) satellites, Sentinel-1A and Sentinel-1B, are also part of the Copernicus mission and were launched in 2014 and 2016 (Torres et al., 2012). Sentinel-1 (S1) and Sentinel-2 (S2) images are available with high temporal and spatial resolutions, which is well suited for precision agriculture. S2 images have been widely used for crop mapping and monitoring, *e.g.*, for the detection of land use anomalies (Kanjir et al., 2018), the estimation of biophysical parameters such as the leaf area index (LAI) (Verrelst et al., 2015; Albughdadi et al., 2021) or more generally to provide information on the vegetation status (Defourny et al., 2019). Since S1 images can provide information regarding water content and the structure of the vegetation (Khabbazan et al., 2019), they have been also largely investigated for crop monitoring and mapping (Vreugdenhil et al., 2018; Mansaray et al., 2020; Nasirzadehdizaji et al., 2021). Finally, the joint use of these sensors has been motivated by their complementary (Inglada et al., 2016; Navarro et al., 2016; Veloso et al., 2017; Denize et al., 2018; Mouret et al., 2021).

A main challenge that can affect all the aforementioned applications is the presence of missing data, which is an inherent problem in remote sensing. Multispectral images are particularly sensitive to this issue since they are affected by clouds (to a lesser extent, acquisition problems can also affect SAR images). An illustrative example is provided in Figure 1, where part of a S2 image is covered by clouds. The problem of missing data is of crucial importance when using machine learning techniques, which generally assume a complete feature matrix. The lack of timely information on crops has been identified for decades as a main limitation for precision agriculture based on remote sensing (Moran et al., 1997). This paper focuses on the reconstruction of multiple parcel-level features extracted from S1 and S2 data, when part of S2 images are missing due to the presence of clouds. The features used in this study were previously investigated for the detection of abnormal crop development at the parcel level (Mouret et al., 2021), after discarding the parcels with missing values, which is obviously not acceptable for operational applications. To that extent, we also propose to evaluate the interest of considering partially cloudy images (with missing data) for this application.

Figure 1: S2 image (natural color) acquired in May 09 2018 (tile T31UCP analyzed in this paper). A part of the image is covered by clouds (bottom right)

Various methods have been proposed in the remote sensing literature to deal with missing data. A general review (Shen et al., 2015) has grouped the different methods into four categories: 1) spatial-based methods, 2) spectral-based methods, 3) temporal-based methods and 4) hybrid methods (combining the spatial, spectral and temporal strategies). For crop monitoring, temporal-based and hybrid methods are generally used, since the temporal information is an essential indicator when analyzing the vegetation status. Temporal-based methods are also known as “gap filling” and traditionally rely on linear or spline interpolations. They are well suited to dense noisy time series and have provided

interesting results, *e.g.*, for the classification of crop types or the prediction of plant diversity (Inglada et al., 2015; Vuolo et al., 2017; Fauvel et al., 2020). However, they can lack precision when there is a need to monitor abrupt changes or when data from a large period of time is missing. Hybrid methods have been used intensively in remote sensing, mostly because they are able to impute missing data in multimodal signals and images, such as multispectral and SAR images. Recent techniques based on deep learning have also been investigated for SAR-Optical image matching (Mazza et al., 2018; Hughes et al., 2019). Image matching can be interesting to reconstruct large parts of S2 image. However it generally uses a single SAR image acquired at a date close to the multispectral image to be reconstructed. Consequently, this method does not fully exploit all the available data acquired throughout the growing season. To address this issue, some methods have been designed to reconstruct S2 images based on multi-temporal SAR/Optical data (Bermudez et al., 2019; Ebel et al., 2022). However, these methods are not suited to our use case for various reasons: 1) they require a supervised learning step whereas the proposed strategy is fully unsupervised, avoiding any costly labelling, 2) they focus on the reconstruction of full S2 images while in this study we focus on parcel-level statistics of vegetation indices and 3) they necessitate a huge amount of training data (for example, Ebel et al. (2022) uses a training set composed of 40 ROI of size 4000×4000 px^2). Deep learning methods have also been used to regress NDVI time series based on SAR times series and various other external indicators (e.g., weather, terrain) (Garioud et al., 2020, 2021). However, while being relevant to impute dense time series for large scale applications, these methods need a huge amount of training data (more than 23850 parcels are analyzed in Garioud et al. (2020) and even more in Garioud et al. (2021)), which is not always accessible in practice. For instance, the French Land Parcel Identification System (LPIS) used in these studies is generally available with a delay of one or two years. Moreover, the method proposed by Garioud et al. (2021) has been designed to express NDVI as a function of SAR time series and does not exploit the available S2 information for the imputation task. Similarly, Pipia et al. (2019) proposed to estimate the leaf area index (LAI) at the pixel level using Gaussian processes. Nevertheless, this method has been designed to reconstruct a single optical time series using a single SAR time series, which is too restrictive for the problem addressed here. Finally, interesting results have been obtained with methods ignoring missing data induced by clouds for specific applications, such as land cover classification (Rußwurm and Körner, 2017; Crisóstomo de Castro Filho et al., 2020). However, this paper focuses on a more generic use case, since having access to reconstructed time series can be interesting for a wider range of applications.

The method investigated in this paper can impute missing features derived from S2 data in a robust fashion by using a Gaussian Mixture Model (GMM) (Dempster et al., 1977) learned using the Expectation-Maximization (EM) algorithm. The main originality of the proposed approach is to use outlier scores resulting from an outlier detection algorithm within the EM algorithm to 1) detect abnormal agricultural parcels and 2) have a robust parameter estimation of the GMM parameters. GMMs have been used successfully in remote sensing, *e.g.*, for clustering (Lopes et al., 2017) and supervised classification (Tadjudin and Landgrebe, 2000; Lagrange et al., 2017). However, despite their natural ability to reconstruct missing data (Ghahramani and Jordan, 1994; Eirola et al., 2014), they have not been investigated for crop monitoring (to the best of our knowledge). The main motivation for using GMMs is their faculty to learn complex behaviors in a fully unsupervised way. Even if these models also suffer from the curse of dimensionality, they can be used with a limited amount of data, which is important here since the number of analyzed parcels is relatively small (the database used in the experiments contains around 2000 parcels).

The rest of this paper is organized as follows. Section 2 presents the study area with the available data and the proposed method. Experimental results are presented in Section 3, with an application to the detection of anomalies in the development of rapeseed crops (additional experiments were conducted on wheat crops and are available in the supplementary materials). In Section 4, a discussion on the results and the different imputation methods is proposed. Finally, some conclusions are drawn in Section 5.

2. Materials and Methods

2.1. Data

2.1.1. Study Area and Parcels

The area of interest is located in France (Beauce region) and corresponds to the S2 tile T31UCP (whose center is located approximately at 48°24'N latitude and 1°00'E longitude) as depicted in Figure 2. In this study area, 2218 rapeseed parcels are monitored for the 2017/2018 growing seasons. All the parcels affected by clouds were discarded in a first analysis in order to have a reliable ground-truth to validate the proposed imputation method. Note that experiments were also conducted on 3361 wheat parcels (for the growing season 2016/2017). Since the results

obtained confirmed those obtained on rapeseed parcels, they are reported in the supplementary material attached to this document for the sake of conciseness.

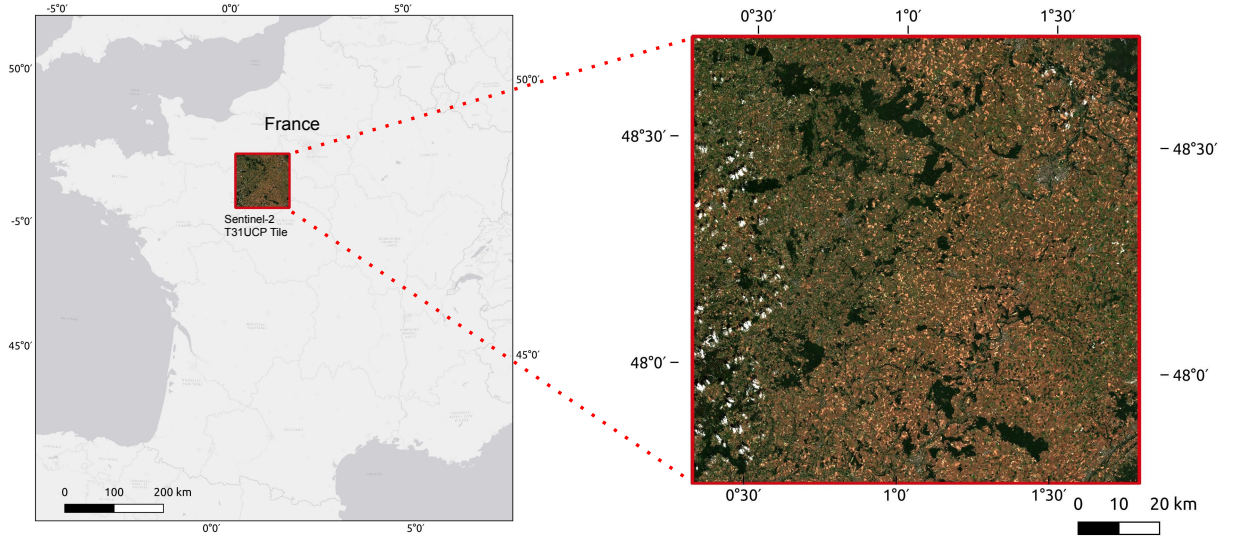


Figure 2: The Sentinel-2 tile considered in this work is delimited by the red box. The S2 image processed in level 2A acquired in June 28 2018 is displayed in natural colors in the right part of the Figure.

2.1.2. Satellite Data

S1-A and S1-B satellites are SAR C-band imaging satellites whose center frequency is 5.405 GHz. The data collected for this study has been acquired in both ascending and descending orbit passes and in dual polarization (VV and VH). Ground Range Detected (GRD) products are used in the Interferometric Wide (IW) swath mode (phase information is lost but the volume of data is drastically reduced) (Torres et al., 2012). The following preprocessing steps have been applied to each S1 image: thermal noise removal, calibration, terrain flattening and range Doppler terrain correction. Multi-temporal speckle filtering was tested, but without bringing any improvement for the considered application. Overall, 40 S1 images were selected for the growing season analyzed. Note that a drop in the number of available S1 images was observed after April 2018 (this can be observed in all web pages related to sentinel data, confirming an acquisition problem).

S2-A and S2-B satellites are multispectral imaging satellites with 13 spectral bands covering the visible, the near infra-red (NIR) and the shortwave-infrared (SWIR) spectral region (Drusch et al., 2012). Level-2A ortho-rectified products expressed in surface reflectance are obtained using the MAJA processing chain (Hagolle et al., 2015), which also provides cloud and shadow masks. Overall, 13 S2 images were selected for the considered growing season. This low number of images is due to the cloud coverage, which often leads to images fully covered by clouds, especially in winter. For instance, no S2 image is exploitable between December 2017 and the end of February 2018, as shown in Figure 3.

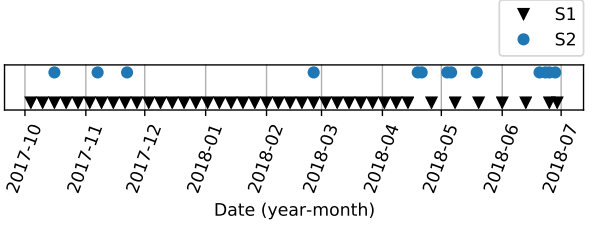


Figure 3: Each marker corresponds to the acquisition date of a used image for the growing season 2017/2018.

Table 1
Pixel-level features extracted from S2 and S1 images used for the crop analysis. For S2, the near infrared (band 8), red edge (band 5), short wave infrared (band 11), green (band 3) and red (band 4) channels are denoted as NIR, RE, SWIR, GREEN and RED, respectively (Mouret et al., 2021). S2 indicators are in the range [-1,1], except for MCARI/OSAVI varying in the range $[0, +\infty]$.

Sensor type	Indicator	Expression
Multispectral	NDVI	$(\text{NIR} - \text{RED})/(\text{NIR} + \text{RED})$
	$\text{NDWI}_{\text{SWIR}}$	$(\text{NIR} - \text{SWIR})/(\text{NIR} + \text{SWIR})$
	$\text{NDWI}_{\text{GREEN}}$	$(\text{GREEN} - \text{NIR})/(\text{GREEN} + \text{NIR})$
	MCARI/OSAVI	$[(\text{RE} - \text{IR}) - 0.2(\text{RE} - \text{RED})] \,/\, \left[(1 + 0.16)\frac{\text{NIR}-\text{RED}}{\text{NIR}+\text{RED}+0.16}\right]$
	GRVI	$(\text{GREEN} - \text{RED})/(\text{GREEN} + \text{RED})$
SAR	Cross-polarized backscattering coefficient VH	γ^0_{VH}
	Co-polarized backscattering coefficient VV	γ^0_{VV}

2.1.3. Pixel-level features

The features extracted at the pixel-level from S1 and S2 images are detailed in Table 1. Each of these features showed interest for crop monitoring and the detection of anomalous development, as detailed in (Mouret et al., 2021). S2 features consist of 5 benchmark agronomic vegetation indices (VI), namely the Normalized Difference Vegetation Index (NDVI) (Rouse et al., 1974), the Green-Red Vegetation Index (GRVI) (Motohka et al., 2010), two variants of the Normalized Difference Water Index (NDWI) (Gao, 1996; McFeeters, 1996) and a variant of the Modified Chlorophyll Absorption Ratio Index (MCARI/OSAVI) (Daughtry et al., 2000; Wu et al., 2008). These 5 VI use different spectral bands combinations, each one focusing on a particular spectral response of the vegetation. S1 features are directly the VV and VH backscattering coefficients. Combinations of these coefficients were tested (e.g., their ratio) but without improving the quality of the detection results (Mouret et al., 2021). Note that other features could be added or removed (depending on the application) without changes in the proposed imputation approach.

2.1.4. Parcel-level features and creation of the feature matrix

Parcel-level features are computed from pixel-level features by using two spatial statistics: the median and the interquartile range (IQR). The median is a robust mean value of an indicator, and can capture the mean behavior of a given parcel. The IQR captures the dispersion of an indicator, providing relevant information at the parcel-level regarding potential heterogeneous development of the vegetation. Note that IQR is not computed for the S1 features, since it is directly proportional to the median. Each parcel is characterized by the different parcel-level statistics computed at each acquisition. The number of columns of the feature matrix is $N_c = N_{1,i} \times N_{1,f} \times N_{1,s} + N_{2,i} \times N_{2,f} \times N_{2,s}$, where $N_{1,i}$ is the number of S1 images, $N_{1,f}$ is the number of pixel-level features extracted for each S1 image, $N_{1,s}$ is the number of statistics computed for each S1 feature. Similar definitions apply to $N_{2,i}$, $N_{2,f}$ and $N_{2,s}$ for S2 images. The processing chain that leads to the creation of the feature matrix used for the detection of outlier parcels is summarized in Figure 4. Each line of the feature matrix has 210 features characterizing each rapeseed parcel (when considering all the acquisition dates and all the spatial statistics, i.e., 130 S2 features and 80 S1 features).

Note that the study area and the data were also used in (Mouret et al., 2021). The interested reader is invited to consult this reference for complementary information and additional examples, especially regarding the description of the parcels with anomalous development.

2.2. Imputation of Missing Values with Mixture of Gaussians using the EM Algorithm

The proposed approach uses a multivariate GMM to impute the potential missing values of the feature matrix. The GMM parameters can be estimated in a fully unsupervised manner, using the Expectation-Maximization (EM)

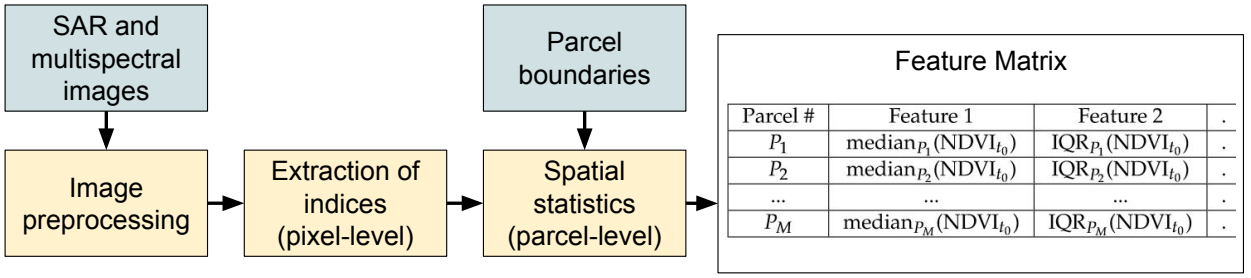


Figure 4: The different steps leading to the creation of the feature matrix used for the detection of parcels with anomalous behavior. NDVI_{t_n} denotes the NDVI computed at time instant t_n while median_{P_M} and IQR_{P_M} denote the spatial median and IQR for the indicator computed inside the parcel $\#\boldsymbol{M}$.

algorithm. Within the EM algorithm, maximum likelihood (ML) estimation of the model parameters is conducted. This estimation can be naturally conducted in the presence of missing data (Dempster et al., 1977; Ghahramani and Jordan, 1994). In that case, the ML estimation is evaluated on the observed values and missing values are estimated using the current model parameters. The resulting probabilistic model is very flexible and can be applied to data of relative small size. After presenting the general principle of the EM algorithm for GMM estimation, we introduce a robust modification of this method taking into account the presence of outliers in the dataset and improving the estimation of the model parameters. More details about GMMs can be found in the classic book from Bishop (2006), while an interesting review dealing with regularization techniques for GMM in high dimension was proposed in Bouveyron and Brunet-Saumard (2014). Regarding the different application of GMMs for agricultural machine vision systems, it is worth mentioning the review proposed by Rehman et al. (2019). An interesting application of the EM algorithm for the detection of cucumber disease was considered in Zhang et al. (2017). In what follows, we only provide key results for the estimation of GMM with the EM algorithm in the presence of missing data. More precisely, Section2.2.1 recalls the principles of the EM algorithm for imputing missing data using GMM, which has been studied for instance in Dempster et al. (1977); Ghahramani and Jordan (1994). Section2.2.2 explains how to robustify this algorithm to mitigate the presence of outliers, which is the main contribution of this work. Detailed derivations and theoretical justifications are provided in Appendix A.1. The algorithm used for the proposed robust EM for GMM is also provided in Appendix A.3.

2.2.1. Imputation of missing data using the EM algorithm for GMM

In what follows, we suppose that each of the N samples of the feature matrix X is distributed according to a mixture of K Gaussian distributions. In the presence of missing values, each sample can be decomposed into $\boldsymbol{x}_n = (\boldsymbol{x}_n^{o_n}, \boldsymbol{x}_n^{m_n})$, where $\boldsymbol{x}_n^{o_n}$ and $\boldsymbol{x}_n^{m_n}$ are the vectors of observed and missing variables respectively. More generally, the superscripts o_n and m_n denote the observed and missing components of the sample n. These subscripts can be used for matrices too, *e.g.*, $\boldsymbol{\Sigma}_k^{o_n m_n}$ refers to the elements of the matrix $\boldsymbol{\Sigma}_k$ in the rows and columns specified by o_n and m_n (and so on). For brevity, we will denote $o_n = o$ and $m_n = m$. The EM algorithm aims at maximizing the observed (or complete) log-likelihood $\log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}^o, \boldsymbol{X}^m, \boldsymbol{z})$ with respect to the model parameters:

$$\log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}^o, \boldsymbol{X}^m, \boldsymbol{z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log \left[\pi_k \mathcal{N}(\boldsymbol{x}_n^o | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right], \tag{1}$$

with $\boldsymbol{X}^o$ the set of all observed variables, $\boldsymbol{X}^m$ the set of all missing variables, $\mathcal{N}(\boldsymbol{x}_n^o | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ the marginal multivariate normal probability density of the observed sample $\boldsymbol{x}_n^o$ and $\boldsymbol{\theta} = \{\pi_1, ..., \pi_k, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K\}$ the set of parameters to be estimated. These parameters are the mean vectors $\boldsymbol{\mu}_k$, the covariance matrices $\boldsymbol{\Sigma}_k$ and the mixing coefficients π_k ($0 < \pi_k < 1, \sum_{k=1}^{K} \pi_k = 1$) of the GMM.

The E-step evaluates $E[\log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}^o, \boldsymbol{X}^m, \boldsymbol{z})|\boldsymbol{\theta}, \boldsymbol{x}^o]$. In practice, it reduces to compute the following sufficient statistics:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n^o, \boldsymbol{\mu}_k^o, \boldsymbol{\Sigma}_k^{oo})}{\sum_{j=1}^K \pi_j \mathcal{N}(\boldsymbol{x}_n^o, \boldsymbol{\mu}_j^o, \boldsymbol{\Sigma}_j^{oo})}, \tag{2}$$

$$\hat{\boldsymbol{\mu}}_{nk}^m = \boldsymbol{\mu}_k^m + \boldsymbol{\Sigma}_k^{mo}(\boldsymbol{\Sigma}_k^{oo})^{-1}(\boldsymbol{x}_n^o - \boldsymbol{\mu}_k^o), \tag{3}$$

$$\hat{\boldsymbol{x}}_{nk}^m = (\boldsymbol{x}_n^o, \boldsymbol{\mu}_{nk}^m), \tag{4}$$

$$\hat{\boldsymbol{\Sigma}}_{nk}^{mm} = \boldsymbol{\Sigma}_k^{mm} - \boldsymbol{\Sigma}_k^{mo}(\boldsymbol{\Sigma}_k^{oo})^{-1}\boldsymbol{\Sigma}_k^{mo}, \tag{5}$$

$$\hat{\boldsymbol{\Sigma}}_{nk} = \begin{pmatrix} \mathbf{0}^{oo} & \mathbf{0}^{om} \\ \mathbf{0}^{mo} & \hat{\boldsymbol{\Sigma}}_{nk}^{mm} \end{pmatrix}, \tag{6}$$

where $\mathbf{0}^{oo}, \mathbf{0}^{om}$ and $\mathbf{0}^{mo}$ are matrices of zeros of appropriate dimensions.

The responsibility γ_{nk} is the probability that samples n has been generated by the k-th component. More precisely, $\gamma_{nk} = E[z_{ik}|\boldsymbol{\theta}^{(t)}, \boldsymbol{x}^o]$ (this is similar to the complete-data case except that γ_{nk} is evaluated on the observed part of the data). The other terms are specific to the missing data case. They results from the evaluation of $E\left[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)|\boldsymbol{\theta}, \boldsymbol{x}_n^o\right]$. Note that (3) and (5) are the conditional expectation and the conditional covariance matrix of the missing variables of a sample given that $\boldsymbol{x}_n$ has been generated by Gaussian #k, *i.e.*, $\hat{\boldsymbol{\mu}}_{nk}^m = E[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]$ and $\hat{\boldsymbol{\Sigma}}_{nk}^{mm} = \text{Var}[\boldsymbol{x}_n^m|\boldsymbol{x}_n^o]$. Note also that the missing values of x_{nk} have been replaced by their expectations $\hat{\boldsymbol{\mu}}_{nk}^m$ in (4). Similarly, the matrix $\hat{\boldsymbol{\Sigma}}_{nk}$ of (6) has been filled with zeros except for the missing components, which leads to $\hat{\boldsymbol{\Sigma}}_{nk}^{mm}$.

In the M-step, maximizing the current expectation lead to the following new set of parameters:

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N_k}\sum_{n=1}^N \gamma_{nk}\hat{\mathbf{x}}_n,\ \boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N_k}\sum_{n=1}^N \gamma_{nk}\left[(\hat{\mathbf{x}}_n - \hat{\boldsymbol{\mu}}_k)(\hat{\mathbf{x}}_n - \hat{\boldsymbol{\mu}}_k)^T + \hat{\boldsymbol{\Sigma}}_{nk}\right],\ \pi_k^{(t+1)} = \frac{N_k}{N}, \tag{7}$$

where with $N_k = \sum_{n=1}^N \gamma_{nk}$. These expressions are similar to the complete-data case, except that missing values have been imputed using the expression obtained in the E-step and covariance matrices are corrected to take into account the presence of missing values. More details about the EM algorithm for the GMM with missing data can be found for instance in Ghahramani and Jordan (1994) and it is worth mentioning the interesting work conducted in Eirola et al. (2014) devoted to the estimation of distances with missing values and applied to various tasks including classification and regression.

Finally, note that when estimating GMMs, a regularization of the covariance matrices is generally needed to avoid instabilities and numerical issues. The strategy adopted for our implementation is based on the method proposed by Bouveyron et al. (2007). It consists in setting the smallest eigenvalues of the covariance matrices to a constant. More details on that point are available in Appendix A.2.

2.2.2. Robust GMM

The estimation of the means and covariances of a GMM using the EM algorithm is sensitive to the presence of outliers, especially in the M-step (Campbell, 1984; Tadjudin and Landgrebe, 2000). To address this issue in the context of semi supervised classification with remote sensing images, Tadjudin and Landgrebe (2000) introduced a robust parameter estimation method which associates weights to the observed samples. The idea is that samples with a reduced weight (corresponding ideally to outliers) will have a small influence on the estimation of the model parameters. However, the method proposed in Tadjudin and Landgrebe (2000) suffers from two main limitations: 1) It does not detect the outliers in an unsupervised way and 2) It does not take into account the presence of missing data. To overcome these issues, we propose to modify this method by using the output of the Isolation Forest (IF) algorithm, which is a reference method for the detection of the outliers (Liu et al., 2012). This algorithm was found to be efficient to detect relevant abnormal parcels (Mouret et al., 2021) and has the advantage of providing an outlier score in the range [0,1]. In order to build a robust GMM, we propose to weight the importance of each sample in the M-step by using the anomaly score provided by the IF algorithm. The resulting robust EM algorithm updates the unknown GMM

parameters in the M-step as in Tadjudin and Landgrebe (2000)

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{n=1}^{N} w_n \gamma_{nk} \hat{\mathbf{x}}_{nk}}{\sum_{n=1}^{N} w_n \gamma_{nk}}, \boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{N} w_n^2 \gamma_{nk} \left[(\hat{\mathbf{x}}_n - \hat{\boldsymbol{\mu}}_k)(\hat{\mathbf{x}}_n - \hat{\boldsymbol{\mu}}_k)^T + \hat{\boldsymbol{\Sigma}}_{nk} \right]}{\sum_{n=1}^{N} w_n^2 \gamma_{nk}}, \pi_k^{(t+1)} = \frac{N_k}{N}. \tag{8}$$

However, contrary to Tadjudin and Landgrebe (2000), the weights w_n are computed using the outlier score of the IF algorithm (denoted as $\text{score}_{\text{IF}}(\hat{\mathbf{x}}_n)$ for the imputed sample $\hat{\mathbf{x}}_n$) as follows:

$$w_n = \frac{1}{1 + \exp\left[\alpha(\text{score}_{\text{IF}}(\hat{\mathbf{x}}_n) - \text{th})\right]}, \tag{9}$$

where α and th are two constants to be fixed by the user. Motivations for using the sigmoid (9) include the fact that it is a smooth monotonically function of the weights taking its values in the range [0,1], with a unique inflection point equal to th. Note that the parameter α controls the speed of the inflection: for high values of th, the sigmoid (9) reduces to a hard thresholding operation around th, whereas it decreases more slowly from 1 to 0 for lower values of th. A score of 0.5 is a natural threshold in the IF algorithm, as explained in Liu et al. (2012). An example of the evolution of the weights with respect to their anomaly score is depicted in Figure 5 for $\alpha = 50$ and $\text{th} = 0.5$.

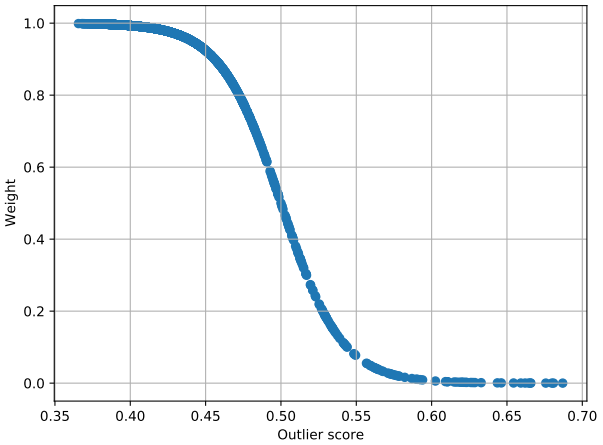


Figure 5: Variation of the weight w_n versus the outlier score attributed by the IF algorithm, with $\alpha = 50$ and $\text{th} = 0.5$.

2.3. Evaluation method and simulation scenarios

This section provides details on the simulation scenarios, which were designed to evaluate the presence of missing values in the observed data. We also introduce metrics to evaluate the performance of the different algorithms considered to reconstruct the missing data. An application to the detection of anomalies in agricultural parcels is finally described with appropriate performance measures.

2.3.1. Simulation scenarios

In order to evaluate the performance of missing data reconstruction with a controlled ground truth, we removed some existing features in the dataset introduced in Section 2.1.4. Two parameters control the number of missing data: the percentage of S2 images having missing values (*e.g.*, due to the presence of clouds), and for each of these S2 images, the percentage of parcels affected by missing values (the parcels affected by clouds are not necessarily the same for each S2 image). For a given S2 image with missing data, we have removed all the features associated with this image for all the affected parcels. In practice, for cloudy days, missing values are likely to affect a significant amount of the parcels. In the presented experiments, half of the total number of parcels (chosen equally likely in the database) was supposed to be cloudy with all S2 features removed (other tests were made with different percentages of cloudy images leading to similar conclusions). The different scenarios considered in this paper are summarized in Table 2.

Table 2
Summary of the experiments conducted in this paper with the percentage of cloudy S2 images and the percentage of cloudy parcels within a given cloudy S2 image. The column "Cloudy S2 images" indicates the percentage of S2 images with missing values whereas the column "Affected parcels" provides the percentage of parcels with missing values within a cloudy S2 image.

Section	Evaluated factor	Cloudy S2 images	Affected parcels
B	Convergence of the EM algorithm	8% (1 S2 image)	50%
3.1.1	Effect of the percentage of S2 images affected by missing values	Varies in [0, 70]%	50%
3.1.2	Effect of adding irrelevant samples	23% (3 S2 images)	50%
3.2	Application to the detection of abnormal crop development	Varies in [0, 70]%	50%

2.3.2. Performance measures

The mean absolute reconstruction error (MAE) is used to evaluate the quality of the reconstruction of the different missing features, with the advantage of being unambiguous and naturally understandable compared to the root mean squared error (RMSE) (Willmott and Matsuura, 2005). The MAE is defined as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{N_m} |f_i - \hat{f}_i|}{N_m}, \tag{10}$$

where N_m is the number of missing features, f_i is the original value of the ith feature and $\hat{f}_i$ denotes its estimation (also referred to as imputation or reconstruction).

The second set of experiments shows the usefulness of the proposed method for the detection of abnormal crop development in the presence of missing data. Using an outlier detection algorithm (here, the IF algorithm) to detect potential anomalies, the precision of the results (defined by the number of true positives divided by the total number of parcels detected as outliers) is computed for various outlier ratios. This outlier ratio is fixed by the user and corresponds to the percentage of abnormal parcels to be detected. The area under the precision vs. outlier ratio curve (AUC) is a good metric to measure the ability of an outlier detection algorithm to detect relevant anomalies. It is similar to the precision vs. recall curves or receiver operational characteristics with the advantage of being adapted to imbalanced datasets containing outliers (Saito and Rehmsmeier, 2015; Mouret et al., 2021).

3. Results

Both robust and non-robust GMM imputation methods are compared to the imputation obtained using the k-nearest neighbors (KNN)(Troyanskaya et al., 2001). Note that the non-robust version of the GMM is regularized using the technique mentioned in Section A.2. Various other imputation methods (gap filling, autoencoders, multiple imputations, soft imputation) were tested and are discussed in Section 4. They are not presented here for conciseness since they did not improve our results. The results presented in this paper focus on the imputation of multispectral S2 time series, but the same method could be used to reconstruct S1 features as well. Finally, note that before the GMM and KNN imputations, each feature was scaled in the range $[0, 1]$ using the available values, *i.e.*, without the missing data (features in natural scale can then be obtained using the inverse transformation). The hyperparameters used for the different reconstruction algorithms are reported in Table 3, with more details regarding the convergence and parameter selection included in Appendix B.

Table 3
Hyperparameters used in the experiments for the GMM and KNN algorithms. R-GMM refers to robust GMM.

Algorithm	Hyperparameter	Values
GMM	K	Estimated using BIC
GMM	Regularization parameter (scree test)	10^{-5}
R-GMM	th	0.5
R-GMM	α	40
KNN	Number of neighbors k	5

3.1. Data imputation

This section evaluates the imputation performance of the proposed GMM method when applied to time series of vegetation indices for crop monitoring. In particular, we test the robustness of the proposed approach to the amount of missing data, and in a second step to the amount of outlier by introducing samples coming from different crop types.

3.1.1. Imputation results obtained by varying the amount of S2 images affected by missing values

The dataset used in this study is relatively exempt of errors coming from the parcel data (e.g, less than 1% of errors in the crop type reported) or the features (e.g., few undetected clouds) as detailed in Mouret et al. (2021). As a consequence, this dataset is a good start to test the imputation methods in controlled conditions.

The influence of the amount of missing data on the imputation results was tested by varying the percentage of S2 images affected by missing values, as depicted in Figure 6. All the results were obtained by averaging the outputs of 50 MC runs. We recall that for each S2 image with missing data, 50% of the parcels were randomly chosen in the database and their corresponding features were removed. The MAE obtained for all the S2 features is depicted in Figure 6(a) whereas Figure 6(b) and (c) show specifically the MAE of the median and IQR NDVI. Note that in Figure 6(a), the S2 features are scaled in the range [0,1] to be able to have comparable results (*e.g.*, MCARI/OSAVI features are not normalized), which can lead to MAE greater than those obtained in natural scale. One can observe that the GMM imputation algorithm outperforms the KNN imputation with good reconstructions even with a high amount of missing data. Results obtained with the classical GMM are close to those obtained with the robust GMM in these experiments. Looking specifically at the median NDVI (Figure 6(b)), it appears that using S1 data is particularly useful, especially when there is a high amount of missing S2 features (the same observation was made for the median statistics of the other VIs studied). Indeed, when the percentage of S2 images with missing data is higher than 40%, one can observe that the MAE is significantly lower when using S1 images (plain curves) to reconstruct median NDVI features (this is also true for S2 features in general, as highlighted in Figure 6(a)). For example, when the percentage of missing S2 images is equal to 70%, the difference in MAE obtained when imputing median NDVI values with and without S1 images is around 0.006 when using GMM imputation methods (corresponding to a reduction of the MAE close to 20%). The reconstruction of IQR statistics is however not favored by the use of S1 data, as shown in Figure 6(c). For this statistics, the robust GMM provides a lower MAE than the classical GMM.

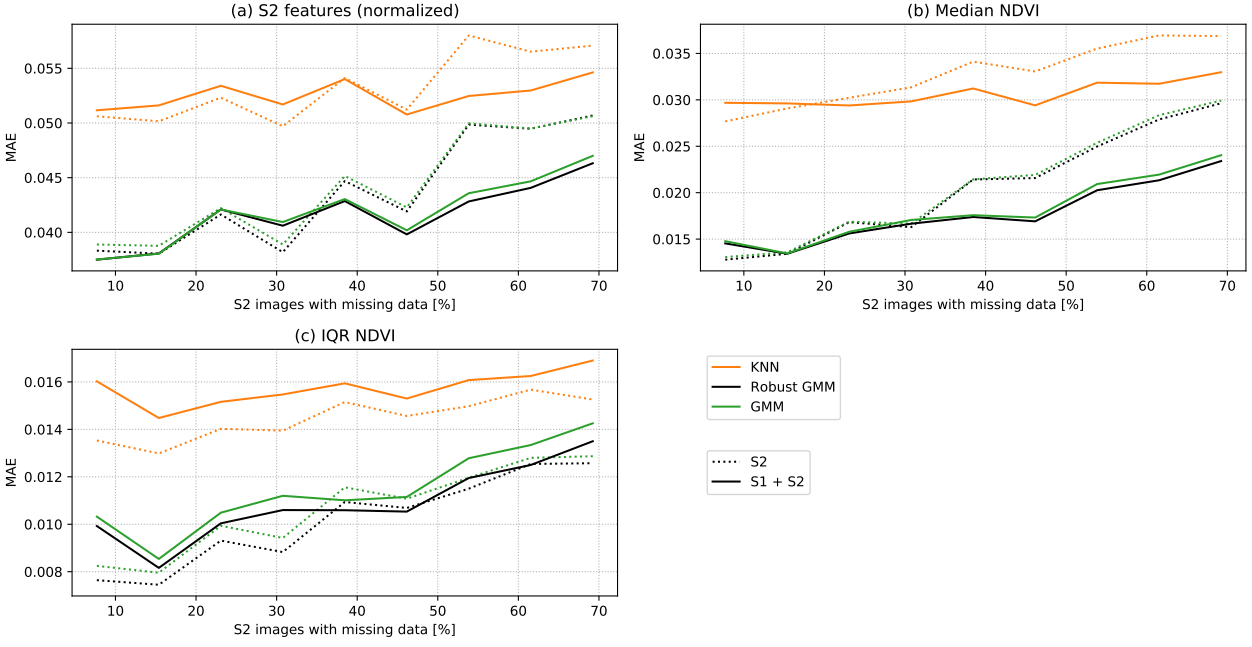


Figure 6: MAE for rapeseed vegetation indices versus the percentage of missing images. X-axis: percentage of S2 images with missing values. Y-axis: MAE for (a) the normalized S2 features (all the S2 indicators are considered), (b) the median of NDVI and (c) the IQR of the NDVI (computed at the parcel level). Results in dotted lines are obtained using S2 features only whereas solid lines correspond to the joint use of S1 and S2 data. The results are averaged after 50 MC runs.

3.1.2. Imputation results obtained by introducing samples coming from different crop types

In practice, errors or noise can contaminate the feature matrix with samples that are very different from the rest of the data. In that case, GMM learning can be more difficult and lead to inaccurate imputations. To investigate the sensitivity of the imputation method to the presence of irrelevant samples, agricultural parcels with a different crop type than rapeseed were included into the rapeseed dataset (these crop types mainly correspond to wheat, maize and barley). The features of these parcels were extracted using field boundaries coming from the French Land Parcel Identification System (LPIS) (Barbottin et al., 2018), which is available in open license.

Imputation results obtained on the rapeseed parcels by varying the percentage of contamination (i.e., the percentage of non-rapeseed parcels in the dataset) are provided in Figure 7, showing the median of the MAE computed using 50 MC runs. The median of the MAE is used here since some extreme MAE values are obtained when using the standard GMM imputation due the presence of non-rapeseed parcels (contrary to the robust GMM). For each run, there are three random S2 images with missing values affecting 50% of the parcels (note that the MAE is computed using the rapeseed parcels only). Using the robust GMM imputation is particularly useful in that case, with an MAE almost stable with respect to the percentage of irrelevant samples in the dataset (it is especially true for the NDVI statistics). Note that the standard GMM imputation is highly impacted by the presence of outliers in the dataset and can lead to large errors, with reconstruction sometimes worse than those obtained using the KNN approach. Consequently, using the robust GMM imputation is recommended in practice, especially if the dataset contains some irrelevant samples.

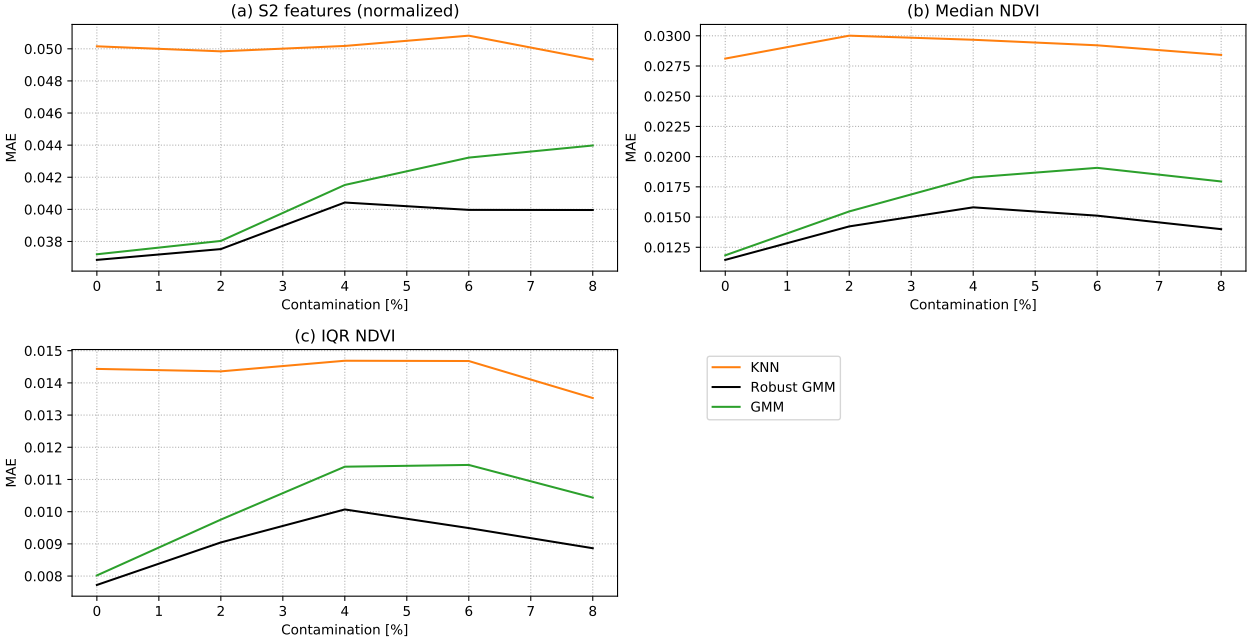
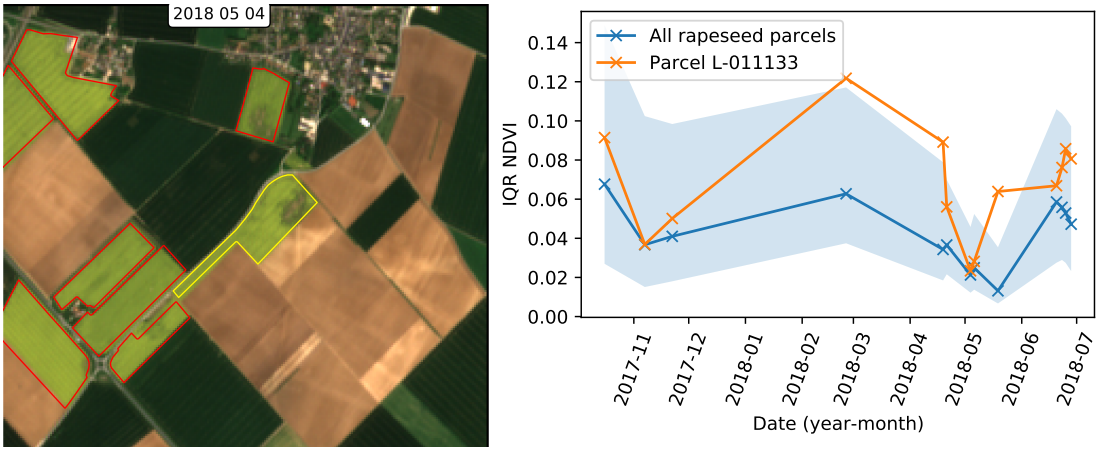


Figure 7: Median of MAE versus the percentage of contamination in the dataset (i.e., coming from non-rapeseed crops) after 50 MC runs for (a) the normalized S2 features (all the S2 indicators are considered), (b) the median of NDVI and (c) the IQR of NDVI (computed at the parcel level). Results are obtained using S1 and S2 features jointly. For each MC run, the percentage of missing data has been fixed: three S2 images (23%) have missing data affecting 50% of the parcels.

3.2. Application to the detection of anomalous crop development in the presence of missing data

Detecting potential anomalies in the development of crop parcels is an important problem in crop monitoring. Such detection can be useful for farmers or agricultural cooperatives to automatically detect parcels with potential problems, without a need for on-site visits. Typical crop anomalies can be grouped into 4 categories: growth anomalies, heterogeneity problems, database problems (*i.e.*, wrong crop type or inaccurate boundaries reported in the database) and false positives (*e.g.*, due to undetected clouds or shadows). An example of heterogeneous parcel is depicted in Figure 8 (a further analysis showed that a part of the parcel was damaged during winter). Each parcel was labeled by an agronomic expert as true positive (relevant anomaly to be detected) or false positive (not relevant for crop monitoring).



(a) (b)

Figure 8: (a) A rapeseed parcel (yellow boundaries) affected by heterogeneity, the image was acquired in May 2018. (b) Interquartile Range (IQR) of the NDVI time series for the yellow parcel (orange line). The blue line is the median value of the whole dataset. The blue area is filled between the 10th and 90th percentiles. The orange line clearly shows an abnormal behaviour of the parcel due here to heterogeneity problems.

Following the method proposed in Mouret et al. (2021), parcels with abnormal behavior are detected using the Isolation Forest algorithm, which computes anomaly scores using the feature matrix (whose construction is detailed in Section 2.1.4). The strongest anomalies are generally related to errors in the crop type (as simulated in Section 3.1.2), whereas other anomalies are parcels with abnormal phenological development. While it was shown that this method is useful to detect relevant anomalies in the crop development, one main problem is that this method cannot be applied in the presence of missing data. As a consequence, we propose to use an imputation method before the outlier detection step, in order to consider images partially covered by clouds. The distribution of the rapeseed parcels detected as abnormal using the IF algorithm with an outlier ratio of 20% is displayed in Figure 9.

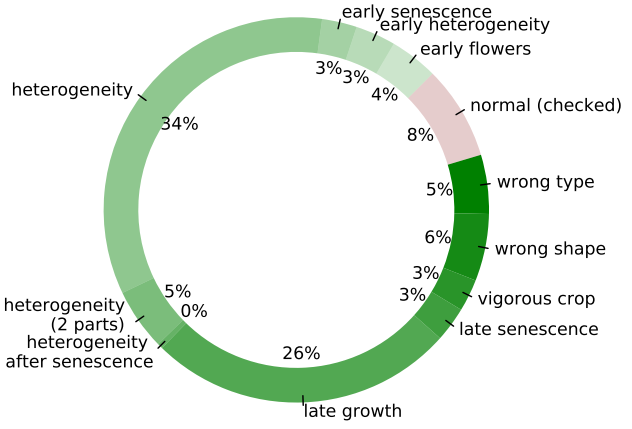


Figure 9: Distribution of the detected rapeseed parcels in each outlier category when using the IF algorithm with an outlier ratio of 20%.

This section evaluates the influence of missing values on the detection results resulting from the application of the IF algorithm. The AUC values (the higher the better) obtained by varying the amount of S2 images affected by missing data are displayed in Figure 10 (more details on this metric were provided in Section 2.3.2). It can be observed that accurate results are obtained with AUC greater than 0.84, even with a high percentage of missing data in the dataset. In particular, the best results are obtained using S1 and S2 data jointly and a reconstruction with the GMM (both robust

and non-robust versions perform similarly in that case). It is interesting to note that discarding S2 images affected by missing values yields a reduced detection performance, since some of the anomalies cannot be detected anymore. Thus, the imputation methods are able to reconstruct the VI with sufficient accuracy to detect the abnormal crop parcels and important information on the parcel behavior seems to be lost without the reconstruction step.

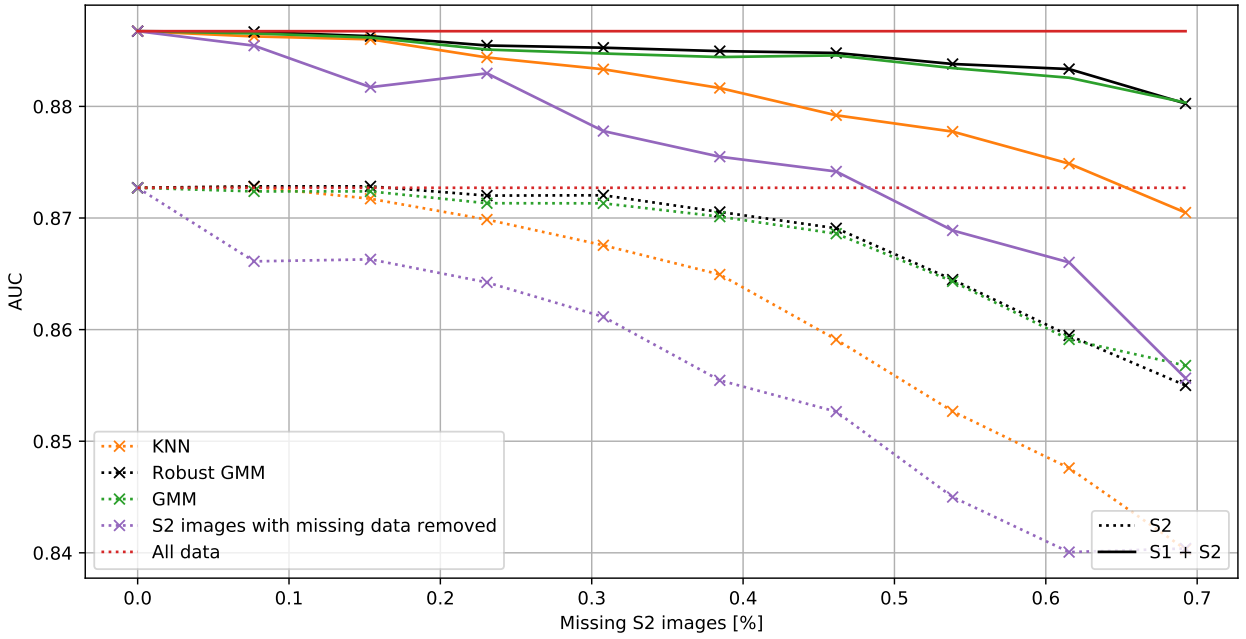


Figure 10: Area under the precision vs. outlier ratio curve (AUC) w.r.t. the percentage of cloudy S2 images (50% of the parcels in a cloudy S2 image have missing data, i.e., do not contain S2 features). Results in dotted lines are obtained using S2 features only whereas solid lines correspond to the joint use of S1 and S2 data. All results are averaged using 50 MC runs.

4. Discussion

This section provides some comments about the results obtained in this paper. Additional experiments (available in the supplementary material) are also discussed.

4.1. Analysis of the presented results

The experiments conducted in this study show that 1) GMM imputations outperform the KNN method 2) using a robust GMM is of crucial importance in the presence of strong outliers. Hence, our results confirm the interest of using outlier detection techniques as standard preprocessing steps in remote sensing, as also recommended for instance in (Pelletier et al., 2017) for the classification of land cover. Note that the obtained results are coherent with the literature: an MAE of 0.0281 was obtained in (Yu et al., 2021) for the reconstruction of NDVI in crop vegetation, an MAE of 0.038 was obtained in (Garioud et al., 2020) for the reconstruction of NDVI for grassland parcels and MAE varying from 0.035 to 0.042 (depending on the region analyzed) was obtained in (Garioud et al., 2021) for agricultural parcels. While these results provide quantitative values for comparison purposes, important differences have to be highlighted: existing studies generally focus on NDVI time series acquired at the pixel-level and do not analyze crops at the parcel level, as proposed in this paper. Moreover, some of these studies focus on the regression of SAR time series towards NDVI (Garioud et al., 2021). Nevertheless, obtaining an MAE close to 0.013 (resp. to 0.019) when imputing the median NDVI for rapeseed (resp. wheat) crops is encouraging (see Tables S1 and S2 in the supplementary materials).

The interest of using a combination of S1 and S2 features was confirmed by our experiments. In particular, using S1 features is interesting to reconstruct more accurately S2 features and thus ensures a better detection of crop anomalies. Two specific examples illustrating the interest of using S1 data are provided in the supplementary material for rapeseed and wheat parcels (Figures S2 and S3). The heterogeneity of a parcel (summarized using IQR at the parcel-level) is

less linked to S1 data, which confirms previous results given in (Mouret et al., 2021). It was also observed that using various features extracted from S2 data helps to reconstruct missing data in NDVI time series when compared to using NDVI only. The interest of using S1 and S2 images jointly has also been confirmed for the detection of anomalous crop development in the presence of missing data, supporting the previous results found in Mouret et al. (2021).

When removing features from one of the S2 image (Figures S4 and S5 in the supplementary material), it appears that some specific stages of the growing season are more difficult to reconstruct, with differences observed for rapeseed and wheat crops. For rapeseed crops, the first S2 acquisition (October 10) is challenging to reconstruct. One explanation is that at this date some fields are not sowed yet whereas others are already vigorous, leading to a higher dispersion of the parcel indicators. The higher MAE obtained for S2 data acquired in February can be explained as follows: 1) S2 images before and after this date correspond to very distant dates 2) crop parcels can be more or less affected by winter, which again leads to a larger dispersion of the indicators. Regarding wheat crops, the high reconstruction errors observed for the data acquired in June 18 can be explained by the beginning of the senescence, which leads to abrupt changes in the crop behavior.

4.2. Other imputation methods

Other strategies for the imputation of missing data were tested without bringing any improvement compared to the proposed method (see Figure S6 in the supplementary material). Some observations are briefly provided below.

Gap filling methods (linear interpolation, spline interpolation and Whittaker smoother) (Cai et al., 2017) perform overall poorly compared to the methods investigated in this paper, which is mainly due to the sparsity of S2 acquisitions, confirming the results found in (Yu et al., 2021). Moreover, when applied to the detection of abnormal crop development, smoothing methods tend to decrease the accuracy of the detection results. Other benchmark imputation methods were tested, such as Multiple Imputation by Chained Equations (MICE) proposed in van Buuren and Groothuis-Oudshoorn (2011) and implemented in the Python library scikit-learn (Pedregosa et al., 2011). Similarly to the KNN imputation, this method provides reconstruction results significantly less accurate than those obtained using the proposed GMM imputation algorithm. Deep learning methods were also tested without success (see an experimental example with Generative Adversarial Networks (Yoon et al., 2018) in Figure S7 of the supplementary materials). The poor results obtained with these methods (which include LSTM, denoising autoencoders (Vincent et al., 2008) or Generative Adversarial Networks (Yoon et al., 2018)) are probably due to the small number of samples in the dataset, which confirms the observations made in Yoon et al. (2018) (Figure 2).

Finally, we considered some outlier detection methods that do not need to impute missing data. It is the case with the IF algorithm, which can be extended to handle missing values without imputation using the strategies studied in (Zemicheal and Dietterich, 2019; Cortes, 2019). This type of strategy is appealing since it drastically reduces the computation time when compared to GMM-based methods. However we observed that these methods are sensitive to the amount of missing values in the dataset and can lead to poor results. Moreover, having access to reliable reconstructed time series is interesting for crop monitoring since it allows the user to analyze with more details the behavior of an abnormal parcel.

4.3. Regularization techniques for GMM

GMM are subject to the curse of dimensionality (Bouveyron and Brunet-Saumard, 2014). This problem was confirmed in our application, especially due to the small number of parcels compared to the high number of features. The regularization of Bouveyron et al. (2007) used in this paper provided the best results overall.

Another classical regularization consists of adding a sparsity constraint to the precision matrices, which can be solved using the graphical lasso algorithm Friedman et al. (2008), which has been adapted to the missing data problem (Ruan et al., 2011). However, using such regularization yielded poor results for the reconstruction of vegetation indices. The sparsity of the precision matrix is due to conditionally independent variables, which is not the case in the proposed feature vector gathering the same features acquired at different time instants. The sparsity of the covariance matrices was also investigated using the method proposed in (Fop et al., 2019) without improving the results obtained with the $[a_{ij}bQ_id_i]$ model suggested in Bouveyron et al. (2007).

5. Conclusion

This paper studied an imputation method based on Gaussian Mixture Models (GMM) for the reconstruction of remote sensing time series constructed from vegetation indices (VI) associated with Sentinel-2 (S2) data. One contribution of this paper is to propose a method able to reconstruct simultaneously various time series, coming from different VI whose statistics have been computed at the parcel-level. These statistics (here, the median and interquartile range) are well suited for crop monitoring since they can characterize efficiently the parcel behaviors, *e.g.*, to detect abnormal growth or heterogeneity problems. This paper showed that using a GMM imputation to reconstruct missing values in the feature matrix performs significantly better than other reference methods such as the k-nearest neighbors or the Multiple Imputation by Chained Equations (MICE)).

Another contribution of this paper is to propose a robust GMM imputation method, which attributes weights to each sample based on the outlier scores resulting from the Isolation Forest algorithm. Samples with high outlier scores have reduced weights, limiting their impact on the estimation of the GMM parameters. Using the proposed robust GMM method instead of the standard GMM imputation method is particularly useful in the presence of irrelevant samples contaminating the dataset. For operational services, we then recommend to use this robust version since it consistently provides reconstruction results similar or better than the standard GMM imputation method.

The experiments conducted in this paper have shown that using Sentinel-1 (S1) features in addition to S2 features improves GMM imputation results in different cases. In particular, using S1 images can be interesting when the amount of missing S2 images is important (above 40%) and for some specific S2 features such as the median NDVI of the parcels. This indicator can be reconstructed with good accuracy (mean absolute error (MAE) close to 0.013 for rapeseed crops and to 0.020 for wheat crops), even with a high amount of missing data (*e.g.*, for rapeseed parcels, the MAE is close to 0.020 even when 70% of the S2 images have 50% of the parcels affected by missing data).

An application to the detection of anomalous crop development in presence of missing data was also investigated. Using S1 and S2 images jointly provided best results for this application. Using a Gaussian mixture model (GMM) for the reconstruction of missing data provided detection results significantly better than with KNN imputation. Note that discarding images affected by clouds leads to poor results for this application.

Further investigations will be conducted to determine whether other regularizations could be applied to GMM to improve the imputation results, for instance by finding an adapted structure for the covariance matrices or by reducing the dimensionality of the dataset. Moreover, since GMM are good models for vegetation indices, other applications, such as forecasting, clustering or classification, would deserve to be investigated. In particular, the automatic classification of the different anomalies could be considered as in León-López et al. (2021). Adding external information such as climate data could also be relevant to reconstruct more efficiently the various VI, since interesting results were obtained for the reconstruction of NDVI time series (Vuolo et al., 2017; Yu et al., 2021). Moreover, adding other type of image data such as unmanned aerial vehicles (UAV) imagery could be also interesting to improve the quality of the imputations (Tang et al., 2021; Huang et al., 2021).

Appendix A The EM algorithm for GMM

A.1 The standard EM algorithm (without missing data)

Given a feature matrix $\boldsymbol{X}$, we assume that each row of this matrix is distributed according to a mixture of K Gaussian distributions. The corresponding log-likelihood can be expressed as (up to an additive constant):

$$\log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right), \tag{A.1}$$

where N is the number of samples in the dataset, $\boldsymbol{x}_n$ is a specific sample, $\mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the probability density function (PDF) of the multivariate normal distribution and $\boldsymbol{\theta} = \{\pi_1, ..., \pi_k, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K\}$ contains the parameters to be estimated. These parameters are the mean vectors $\boldsymbol{\mu}_k$, the covariance matrices $\boldsymbol{\Sigma}_k$ and the mixing coefficients π_k ($0 < \pi_k < 1, \sum_{k=1}^{K} \pi_k = 1$) of the GMM. The maximization of (A.1) with respect to (w.r.t.) $\boldsymbol{\theta}$ being complex, it is classical to introduce labels $\boldsymbol{z} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_n\}$ indicating the Gaussians associated with the observed vectors

$\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ (the maximization of the likelihood is straightforward for known labels) and the complete log-likelihood:

$$\log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log \left[\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right], \tag{A.2}$$

where $z_{nk} = 1$ if the vector $\boldsymbol{x}_n$ belongs to the kth component of the GMM and $z_{nk} = 0$ otherwise. After an appropriate initialization of $\boldsymbol{\theta}$, the EM algorithm aims at maximizing the conditional expectation of $\log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{z})$ in an iterative fashion until convergence. The expectation step (E-step) computes the expectation of the complete log-likelihood conditionally to the current set of the mixture parameters, $\boldsymbol{\theta}^{(t)}$:

$$E[\log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{z}))|\boldsymbol{\theta}^{(t)}] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \log \left[\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right], \tag{A.3}$$

where $\gamma_{nk} = E[z_n = k|\boldsymbol{x}_n, \boldsymbol{\theta}^{(t)}]$ is referred to as responsibilities. The maximization step (M-step) maximizes $E[\log \mathcal{L}_c(\boldsymbol{\theta}; z)|\boldsymbol{\theta}^{(t)}]$ w.r.t. $\boldsymbol{\theta}$ to provide an updated parameter vector $\boldsymbol{\theta}^{(t+1)}$:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} E[\log \mathcal{L}_c(\boldsymbol{\theta}; z)|\boldsymbol{\theta}^{(t)}]. \tag{A.4}$$

For brevity, we will denote $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}$, *i.e.*, $\boldsymbol{\mu}_k^{(t)} = \boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k^{(t)} = \boldsymbol{\Sigma}_k$ and $\pi_k^{(t)} = \pi_k$ the current set of parameters in what follows.

E-step: in practice the E-step reduces to the computation of the responsibilities γ_{nk}, which are also the probabilities that the sample $\boldsymbol{x}_n$ has been generated by the kth Gaussian component:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{A.5}$$

M-step: the parameters are re-estimated using the updated responsibilities:

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n, \; \boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T, \; \pi_k^{(t+1)} = \frac{N_k}{N}, \tag{A.6}$$

with $N_k = \sum_{n=1}^{N} \gamma_{nk}$. The EM algorithm is stopped when the log-likelihood or the parameter values does not change significantly or after a fixed number of iterations.

A.2 Regularization of the covariance matrices

Learning the parameters of a GMM can be subject to instabilities, especially when the covariance matrices are ill-conditioned (in some extreme cases, the covariance matrix cannot even be inverted). A heuristic strategy for regularizing a covariance matrix consists in adding a small constant to its diagonal elements during the estimation (*e.g.*, this regularization is proposed in the Python library scikit learn Pedregosa et al. (2011)). Alternatively, Bouveyron et al. (2007) studied different regularization techniques adapted to the estimation of covariance matrices for high dimensional problems. In this study, we have considered the model referred to as $[a_{ij} b Q_i d_i]$ (see Bouveyron et al. (2007) for details). The idea behind this model is to use an eigendecomposition of the covariance matrices $\boldsymbol{\Sigma}_k = Q_k \Delta_k Q_k^T$ and set the smallest eigenvalues to the same constant $b_k = b$ (which can be justified when the data are obtained in a common acquisition process). This operation significantly reduces the number of model parameters to estimate, which is valuable to fight against the curse of dimensionality. The scree test is used to find the number of eigenvalues to be set to the constant value b (see Bouveyron et al. (2007) for details).

A.3 Proposed algorithm

Algorithm 1 Proposed robust EM algorithm for GMM in the the incomplete data case.

Input: Data $\{\boldsymbol{x}_n\}_{n=1}^N$, K the number of clusters

Output: Clustering labels $\mathcal{Z} = \{z_n\}_{n=1}^N$, parameters $\boldsymbol{\theta} = \{\pi_1, ..., \pi_k, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K\}$ and imputed samples $\hat{\boldsymbol{x}}$

For each sample, identify observed and missing components o and m;
Initialize $\boldsymbol{\theta}^{(0)}$; (*e.g.*, randomly or using the kmeans algorithm)
$t \leftarrow 1$
while not convergence **do**
    for $1 \leq k \leq K$: **do** ▷ **E-step**
        ▷ Compute observed responsibilities:
        $\gamma_{nk}^{(t)} = \frac{\pi_k^{(t-1)} \mathcal{N}(\boldsymbol{x}_n^o, \boldsymbol{\mu}_k^{o(t-1)}, \boldsymbol{\Sigma}_k^{oo(t-1)})}{\sum_{j=1}^K \pi_j^{(t-1)} \mathcal{N}(\boldsymbol{x}_n^o, \boldsymbol{\mu}_j^{o(t-1)}, \boldsymbol{\Sigma}_j^{oo(t-1)})}$
        ▷ Compute conditional expectations:
        $\hat{\boldsymbol{\mu}}_{nk}^{m(t)} = \boldsymbol{\mu}_k^{m(t-1)} + \boldsymbol{\Sigma}_k^{mo(t-1)} (\boldsymbol{\Sigma}_k^{oo(t-1)})^{-1} (\boldsymbol{x}_n^o - \boldsymbol{\mu}_k^{o(t-1)})$
        $\hat{\boldsymbol{\Sigma}}_{nk}^{mm(t)} = \boldsymbol{\Sigma}_k^{mm(t-1)} - \boldsymbol{\Sigma}_k^{mo(t-1)} (\boldsymbol{\Sigma}_k^{oo(t-1)})^{-1} \boldsymbol{\Sigma}_k^{mo(t-1)}$
        Fill in: $\hat{\boldsymbol{x}}_{nk}^{(t)} \leftarrow [\boldsymbol{x}_i^o, \hat{\boldsymbol{\mu}}_{nk}^{m(t)}]$ and $\hat{\boldsymbol{\Sigma}}_{nk}^{(t)} \leftarrow \begin{pmatrix} \mathbf{0}^{oo} & \mathbf{0}^{om} \\ \mathbf{0}^{mo} & \hat{\boldsymbol{\Sigma}}_{nk}^{mm(t)} \end{pmatrix}$
    for $1 \leq k \leq K$: **do** ▷ **M-step**
        $w_n = \frac{1}{1+\exp[\alpha(\text{score}_{\text{IF}}(\hat{\boldsymbol{x}}_n^{(t)}) - \text{th})]}$ ▷ Compute outlier weights using the IF algorithm
        $\pi_k^{(t)} = \frac{1}{N} \sum_{n=1}^N \gamma_{nk}^{(t)}$
        $\boldsymbol{\mu}_k^{(t)} = \frac{\sum_{n=1}^N w_n \gamma_{nk}^{(t)} \hat{\boldsymbol{x}}_{nk}^{(t)}}{\sum_{n=1}^N w_n \gamma_{nk}^{(t)}}$
        $\boldsymbol{\Sigma}_k^{(t)} = \frac{\sum_{i=1}^N w_n^2 \gamma_{nk}^{(t)} \left[(\hat{\boldsymbol{x}}_n^{(t)} - \hat{\boldsymbol{\mu}}_k^{(t)})(\hat{\boldsymbol{x}}_n^{(t)} - \hat{\boldsymbol{\mu}}_k^{(t)})^T + \hat{\boldsymbol{\Sigma}}_{nk}^{(t)} \right]}{\sum_{n=1}^N w_n^2 \gamma_{nk}^{(t)}}$
    $t \leftarrow t + 1$
Set z_i as the index k that has the maximimum p_{ik}.
Final imputation: $\hat{\boldsymbol{x}}_n^m = \sum_{k=1}^K \gamma_{nk}^{(t)} \hat{\boldsymbol{x}}_{nk}^{(t)}$

Appendix B Parameter tuning and convergence

B.1 GMM imputation algorithm

The number of Gaussians K in the GMM was estimated using the Bayesian Information Criterion (BIC) as suggested in Bouveyron and Brunet-Saumard (2014). This estimation provided good results without a need for manually choosing the number of components, which can be difficult in practice (especially for an unsupervised task). For the regularization of the covariance matrix, the stopping criterion of the scree test was set to 10^{-5} (see Bouveyron et al. (2007) for more details on the scree test). We observed that a too small value (typically lower than 10^{-6}) can lead to unstable results whereas too high values (typically 10^{-3}) lead to a deterioration of the imputation results.

The parameters of the weighting function w_n are the threshold th and the slope α. The threshold was fixed to th $= 0.5$, which is a natural value to separate outliers and inliers when using the IF algorithm Liu et al. (2012). The slope parameter was fixed to $\alpha = 40$ by cross validation. Small changes in these parameters did not have a significant impact on the imputation results.

The outputs of the EM algorithm depend on its initialization, which is detailed in what follows. The EM algorithm was initialized by the output of the K-means algorithm with K centroids chosen equally likely in the dataset. This

initialization yields a fast convergence of the EM algorithm obtained in less than 10 iterations. The EM algorithm was stopped when the difference between two consecutive values of the log-likelihood was less than 10^{-3}. In order to analyze the sensitivity of the algorithm to its initialization, we ran 50 Monte Carlo (MC) simulations of the EM algorithm using the same dataset (1 S2 image covered by clouds, 50% of the parcels affected by missing values) with different random initializations and imputed the missing values. The distribution of the MAE obtained for these Monte Carlo runs (evaluated using all the reconstructed features for the parcels with missing data) is displayed in Figure B.1, showing that the values of MAE are very similar, varying in the interval [0.02178, 0.02186]. These results indicate that the EM algorithm is not very sensitive to its initialization for the reconstruction of VI at the parcel level.

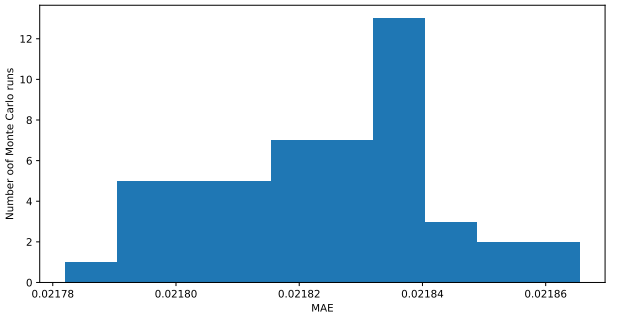


Figure B.1: Histogram of MAE obtained after 50 Monte Carlo runs (with different initializations) on the same dataset.

B.2 KNN imputation algorithm

The KNN imputation method available in the Python library Scikit-Learn (Pedregosa et al., 2011) (version 0.24) (named "KNNimputer") was used as a benchmark. The number of nearest neighbors was fixed to $k = 5$. Changing the value of this parameter in a neighborhood did not have a huge effect on the reconstruction results. The contribution of each neighbor was weighted by the inverse of its distance to the sample to be imputed, similarly to the configuration used in (Albughdadi et al., 2017).

References

Albughdadi, M., Kouamé, D., Rieu, G., Tourneret, J.Y., 2017. Missing data reconstruction and anomaly detection in crop development using agronomic indicators derived from multispectral satellite images, in: Proc. IEEE IGARSS, Fort Worth, TX, USA. pp. 5081–5084. doi:10.1109/IGARSS.2017.8128145.

Albughdadi, M., Rieu, G., Duthoit, S., Alswaitti, M., 2021. Towards a massive Sentinel-2 LAI time-series production using 2-d convolutional networks. Comput. Electron. Agric. 180, 105899. doi:https://doi.org/10.1016/j.compag.2020.105899.

Barbottin, A., Bouty, C., Martin, P., 2018. Using the French LPIS database to highlight farm area dynamics: The case study of the niort plain. Land Use Policy 73, 281 – 289. URL: http://www.sciencedirect.com/science/article/pii/S0264837717302909, doi:https://doi.org/10.1016/j.landusepol.2018.02.012.

Bermudez, J.D., Happ, P.N., Feitosa, R.Q., Oliveira, D.A.B., 2019. Synthesis of multispectral optical images from SAR/Optical multitemporal data using conditional generative adversarial networks. IEEE Geosci. Remote Sens. Lett. 16, 1220–1224. doi:10.1109/LGRS.2019.2894734.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.

Bouveyron, C., Brunet-Saumard, C., 2014. Model-based clustering of high-dimensional data: A review. Comput. Stat. Data Anal. 71, 52–78. URL: https://www.sciencedirect.com/science/article/pii/S0167947312004422, doi:https://doi.org/10.1016/j.csda.2012.12.008.

Bouveyron, C., Girard, S., Schmid, C., 2007. High-dimensional data clustering. Comput. Stat. Data Anal. 52, 502–519. URL: https://www.sciencedirect.com/science/article/pii/S0167947307000692, doi:https://doi.org/10.1016/j.csda.2007.02.009.

van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate imputation by chained equations in R. J. Stat. Softw. 45, 1–67. URL: https://www.jstatsoft.org/v045/i03, doi:10.18637/jss.v045.i03.

Cai, Z., Jönsson, P., Jin, H., Eklundh, L., 2017. Performance of smoothing methods for reconstructing NDVI time-series and estimating vegetation phenology from MODIS data. Remote Sens. 9, 1271. doi:10.3390/rs9121271.

Campbell, N.A., 1984. Mixture models and atypical values. Math. Geol. 16, 465–477. doi:10.1007/BF01886327.

Crisóstomo de Castro Filho, H., Abílio de Carvalho Júnior, O., Ferreira de Carvalho, O.L., Pozzobon de Bem, P., dos Santos de Moura, R., Olino de Albuquerque, A., Rosa Silva, C., Guimarães Ferreira, P.H., Fontes Guimarães, R., Trancoso Gomes, R.A., 2020. Rice crop detection using LSTM, bi-LSTM, and machine learning models from Sentinel-1 time series. Remote Sens. 12, 2655. doi:10.3390/rs12162655.

Cortes, D., 2019. Imputing missing values with unsupervised random trees. arXiv:1911.06646.

Daughtry, C., Walthall, C., Kim, M., de Colstoun, E., McMurtrey, J., 2000. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. Remote Sens. Environ. 74, 229 – 239. doi:https://doi.org/10.1016/S0034-4257(00)00113-9.

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Nicola, L., Rabaute, T., Savinaud, M., Udroiu, C., Valero, S., Bégué, A., Dejoux, J.F., El Harti, A., Ezzahar, J., Kussul, N., Labbassi, K., Lebourgeois, V., Miao, Z., Newby, T., Nyamugama, A., Salh, N., Shelestov, A., Simonneaux, V., Traore, P.S., Traore, S.S., Koetz, B., 2019. Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. Remote Sens. Environ. 221, 551 – 568. URL: http://www.sciencedirect.com/science/article/pii/S0034425718305145, doi:https://doi.org/10.1016/j.rse.2018.11.007.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. 39.

Denize, J., Hubert-Moy, L., Betbeder, J., Corgne, S., Baudry, J., Pottier, E., 2018. Evaluation of using Sentinel-1 and Sentinel-2 time-series to identify winter land use in agricultural landscapes. Remote Sens. 11. URL: https://www.mdpi.com/2072-4292/11/1/37.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens. Environ. 120, 25 – 36. URL: http://www.sciencedirect.com/science/article/pii/S0034425712000636, doi:https://doi.org/10.1016/j.rse.2011.11.026. the Sentinel Missions - New Opportunities for Science.

Ebel, P., Xu, Y., Schmitt, M., Zhu, X., 2022. SEN12MS-CR-TS: A remote sensing data set for multi-modal multi-temporal cloud removal. IEEE Trans. Geosci. Remote Sens., In press doi:https://doi.org/10.48550/arXiv.2201.09613.

Eirola, E., Lendasse, A., Vandewalle, V., Biernacki, C., 2014. Mixture of Gaussians for distance estimation with missing data. Neurocomputing 131, 32–42. URL: https://www.sciencedirect.com/science/article/pii/S0925231213010990, doi:https://doi.org/10.1016/j.neucom.2013.07.050.

Fauvel, M., Lopes, M., Dubo, T., Rivers-Moore, J., Frison, P.L., Gross, N., Ouin, A., 2020. Prediction of plant diversity in grasslands using sentinel-1 and -2 satellite image time series. Remote Sens. Environ. 237, 111536. URL: https://www.sciencedirect.com/science/article/pii/S0034425719305553, doi:https://doi.org/10.1016/j.rse.2019.111536.

Fop, M., Murphy, T.B., Scrucca, L., 2019. Model-based clustering with sparse covariance matrices. Stat. Comput. 29, 791–819. doi:10.1007/s11222-018-9838-y.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9, 432–441. doi:10.1093/biostatistics/kxm045.

Gao, B., 1996. NDWI — A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. 58, 257 – 266. doi:https://doi.org/10.1016/S0034-4257(96)00067-3.

Garioud, A., Valero, S., Giordano, S., Mallet, C., 2020. On the joint exploitation of optical and SAR satellite imagert for grassland monitoring. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLIII-B3-2020, 591–598. URL: https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIII-B3-2020/591/2020/, doi:10.5194/isprs-archives-XLIII-B3-2020-591-2020.

Garioud, A., Valero, S., Giordano, S., Mallet, C., 2021. Recurrent-based regression of sentinel time series for continuous vegetation monitoring. Remote Sens. Environ. 263, 112419. URL: https://www.sciencedirect.com/science/article/pii/S0034425721001371, doi:https://doi.org/10.1016/j.rse.2021.112419.

Ghahramani, Z., Jordan, M., 1994. Supervised learning from incomplete data via an em approach, in: Cowan, J., Tesauro, G., Alspector, J. (Eds.), Advances in Neural Information Processing Systems, Morgan-Kaufmann. pp. 120–127. URL: https://proceedings.neurips.cc/paper/1993/file/f2201f5191c4e92cc5af043eebfd0946-Paper.pdf.

Hagolle, O., Huc, M., Villa Pascual, D., Dedieu, G., 2015. A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of Formosat-2, Landsat, VENμS and Sentinel-2 images. Remote Sens. 7, 2668–2691. URL: https://www.mdpi.com/2072-4292/7/3/2668, doi:10.3390/rs70302668.

Huang, H., Yang, A., Tang, Y., Zhuang, J., Hou, C., Tan, Z., Dananjayan, S., He, Y., Guo, Q., Luo, S., 2021. Deep color calibration for UAV imagery in crop monitoring using semantic style transfer with local to global attention. Int. J. Appl. Earth Obs. Geoinf. 104, 102590. doi:https://doi.org/10.1016/j.jag.2021.102590.

Hughes, L.H., Merkle, N., Bürgmann, T., Auer, S., Schmitt, M., 2019. Deep learning for sar-optical image matching, in: Proc. IEEE IGARSS, Yokohama, Japan. pp. 4877–4880. doi:10.1109/IGARSS.2019.8898635.

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre, G., Bontemps, S., Defourny, P., et al., 2015. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. Remote Sens. 7, 12356–12379. URL: http://dx.doi.org/10.3390/rs70912356, doi:10.3390/rs70912356.

Inglada, J., Vincent, A., Arias, M., Marais-Sicre, C., 2016. Improved early crop type identification by joint use of high temporal resolution SAR and optical image time series. Remote Sens. 8, 362. URL: http://dx.doi.org/10.3390/rs8050362, doi:10.3390/rs8050362.

Kanjir, U., Đurić, N., Veljanovski, T., 2018. Sentinel-2 based temporal detection of agricultural land use anomalies in support of common agricultural policy monitoring. ISPRS Int. J. Geo-inf. 7, 405. doi:10.3390/ijgi7100405.

Khabbazan, S., Vermunt, P., Steele-Dunne, S., Ratering Arntz, L., Marinetti, C., van der Valk, D., Iannini, L., Molijn, R., Westerdijk, K., van der Sande, C., 2019. Crop monitoring using Sentinel-1 data: A case study from the Netherlands. Remote Sens. 11. URL: https://www.mdpi.com/2072-4292/11/16/1887, doi:10.3390/rs11161887.

Lagrange, A., Fauvel, M., Grizonnet, M., 2017. Large-scale feature selection with Gaussian Mixture Models for the classification of high dimensional remote sensing images. IEEE Trans. Comput. Imag. 3, 230–242. doi:10.1109/TCI.2017.2666551.

León-López, K.M., Mouret, F., Tourneret, J.Y., Arguello, H., 2021. Anomaly detection and classification in multispectral time series based on hidden Markov models. In press, IEEE Trans. Geosci. Remote Sens. .

Liu, F.T., Ting, K.M., Zhou, Z.H., 2012. Isolation-based anomaly detection. ACM Trans. Knowl. Discov. Data 6. doi:10.1145/2133360.2133363.

Lopes, M., Fauvel, M., Ouin, A., Girard, S., 2017. Spectro-temporal heterogeneity measures from dense high spatial resolution satellite image time series: Application to grassland species diversity estimation. Remote Sens. 9. URL: https://www.mdpi.com/2072-4292/9/10/993, doi:10.3390/rs9100993.

Mansaray, L.R., Zhang, K., Kanu, A.S., 2020. Dry biomass estimation of paddy rice with Sentinel-1a satellite data using machine learning regression algorithms. Comput. Electron. Agric. 176, 105674. doi:https://doi.org/10.1016/j.compag.2020.105674.

Mazza, A., Gargiulo, M., Scarpa, G., Gaetano, R., 2018. Estimating the NDVI from SAR by convolutional neural networks, in: Proc. IEEE IGARSS, Valencia, Spain. pp. 1954–1957. doi:10.1109/IGARSS.2018.8519459.

McFeeters, S.K., 1996. The use of the normalized difference water index (NDWI) in the delineation of open water features. Int. J. Remote Sens. 17, 1425–1432. doi:10.1080/01431169608948714.

Moran, M., Inoue, Y., Barnes, E., 1997. Opportunities and limitations for image-based remote sensing in precision crop management. Remote Sens. Environ. 61, 319–346. doi:https://doi.org/10.1016/S0034-4257(97)00045-X.

Motohka, T., Nasahara, K.N., Oguma, H., Tsuchida, S., 2010. Applicability of green-red vegetation index for remote sensing of vegetation phenology. Remote Sens. 2, 2369–2387. URL: https://www.mdpi.com/2072-4292/2/10/2369, doi:10.3390/rs2102369.

Mouret, F., Albughdadi, M., Duthoit, S., Kouamé, D., Rieu, G., Tourneret, J.Y., 2021. Outlier detection at the parcel-level in wheat and rapeseed crops using multispectral and SAR time series. Remote Sens. 13, 956. URL: http://dx.doi.org/10.3390/rs13050956, doi:10.3390/rs13050956.

Nasirzadehdizaji, R., Cakir, Z., Balik Sanli, F., Abdikan, S., Pepe, A., Calò, F., 2021. Sentinel-1 interferometric coherence and backscattering analysis for crop monitoring. Comput. Electron. Agric. 185, 106118. URL: https://www.sciencedirect.com/science/article/pii/S0168169921001368, doi:https://doi.org/10.1016/j.compag.2021.106118.

Navarro, A., Rolim, J., Miguel, I., Catalão, J., Silva, J., Painho, M., Vekerdy, Z., 2016. Crop monitoring based on SPOT-5 Take-5 and Sentinel-1A data for the estimation of crop water requirements. Remote Sens. 8, 525. doi:10.3390/rs8060525.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pelletier, C., Valero, S., Inglada, J., Dedieu, G., Champion, N., 2017. New iterative learning strategy to improve classification systems by using outlier detection techniques, in: Proc. IEEE IGARSS, Fort Worth, TX, USA. pp. 3676–3679. doi:10.1109/IGARSS.2017.8127796.

Pipia, L., Muñoz-Marí, J., Amin, E., Belda, S., Camps-Valls, G., Verrelst, J., 2019. Fusing optical and SAR time series for LAI gap filling with multioutput Gaussian processes. Remote Sens. Environ. 235, 111452. URL: https://www.sciencedirect.com/science/article/pii/S0034425719304717, doi:https://doi.org/10.1016/j.rse.2019.111452.

Rehman, T.U., Mahmud, M.S., Chang, Y.K., Jin, J., Shin, J., 2019. Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. Comput. Electron. Agric. 156, 585–605. doi:https://doi.org/10.1016/j.compag.2018.12.006.

Rouse, J., Haas, R., Schell, J., Deering, D., 1974. Monitoring vegetation systems in the great plains with ERTS. NASA special publication 351, 309.

Ruan, L., Yuan, M., Zou, H., 2011. Regularized parameter estimation in high-dimensional gaussian mixture models. Neural. Comput. 23, 1605–1622. doi:10.1162/NECO_a_00128.

Rußwurm, M., Körner, M., 2017. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images, in: IEEE Proc. CVPRW, Honolulu, HI, USA. pp. 1496–1504. doi:10.1109/CVPRW.2017.193.

Saito, T., Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE 10, 1–21. doi:10.1371/journal.pone.0118432.

Schulz, C., Holtgrave, A.K., Kleinschmit, B., 2021. Large-scale winter catch crop monitoring with Sentinel-2 time series and machine learning – an alternative to on-site controls? Comput. Electron. Agric. 186, 106173. URL: https://www.sciencedirect.com/science/article/pii/S0168169921001903, doi:https://doi.org/10.1016/j.compag.2021.106173.

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 Geosci.Remote Sens. Mag. 3, 61–85. doi:10.1109/MGRS.2015.2441912.

Tadjudin, S., Landgrebe, D., 2000. Robust parameter estimation for mixture model. IEEE Trans. Geosci. Remote Sens. 38, 439–445. doi:10.1109/36.823939.

Tang, Y., Dananjayan, S., Hou, C., Guo, Q., Luo, S., He, Y., 2021. A survey on the 5G network and its impact on agriculture: Challenges and opportunities. Comput. Electron. Agric. 180, 105895. doi:https://doi.org/10.1016/j.compag.2020.105895.

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Traver, I.N., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., Huchler, M., Rostan, F., 2012. GMES Sentinel-1 mission. Remote Sens. Environ. 120, 9–24. doi:https://doi.org/10.1016/j.rse.2011.05.028.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays . Bioinformatics 17, 520–525. doi:10.1093/bioinformatics/17.6.520.

Veloso, A., Mermoz, S., Bouvet, A., Toan, T.L., Planells, M., Dejoux, J.F., Ceschia, E., 2017. Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. Remote Sens. Environ. 199, 415 – 426. URL: http://www.sciencedirect.com/science/article/pii/S0034425717303309, doi:https://doi.org/10.1016/j.rse.2017.07.015.

Verrelst, J., Rivera, J.P., Veroustraete, F., Muñoz-Marí, J., Clevers, J.G., Camps-Valls, G., Moreno, J., 2015. Experimental Sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods – a comparison. ISPRS J. Photogramm. Remote Sens. 108, 260 – 272. URL: http://www.sciencedirect.com/science/article/pii/S0924271615001239, doi:https://doi.org/10.1016/j.isprsjprs.2015.04.013.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A., 2008. Extracting and composing robust features with denoising autoencoders, in: Proc. ACM ICML, Helsinki, Finland. p. 1096–1103. URL: https://doi.org/10.1145/1390156.1390294, doi:10.1145/1390156.1390294.

Vreugdenhil, M., Wagner, W., Bauer-Marschallinger, B., Pfeil, I., Teubner, I., Rüdiger, C., Strauss, P., 2018. Sensitivity of Sentinel-1 backscatter to vegetation dynamics: An Austrian case study. Remote Sens. 10, 1396. doi:10.3390/rs10091396.

Vuolo, F., Ng, W.T., Atzberger, C., 2017. Smoothing and gap-filling of high resolution multi-spectral time series: Example of Landsat data. Int. J. Appl. Earth Obs. Geoinf. 57, 202–213. URL: https://www.sciencedirect.com/science/article/pii/S0303243416302100, doi:https://doi.org/10.1016/j.jag.2016.12.012.

Weiss, M., Jacob, F., Duveiller, G., 2020. Remote sensing for agricultural applications: A meta-review. Remote Sens. Environ. 236, 111402. URL: http://www.sciencedirect.com/science/article/pii/S0034425719304213, doi:https://doi.org/10.1016/j.rse.2019.111402.

Willmott, C., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30, 79. doi:10.3354/cr030079.

Wu, C., Niu, Z., Tang, Q., Huang, W., 2008. Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. Agric. For. Meteorol. 148, 1230 – 1241. doi:https://doi.org/10.1016/j.agrformet.2008.03.005.

Yoon, J., Jordon, J., van der Schaar, M., 2018. GAIN: Missing data imputation using generative adversarial nets, in: Proc. PMLR, Stockholm, Sweden. pp. 5689–5698. URL: https://proceedings.mlr.press/v80/yoon18a.html.

Yu, W., Li, J., Liu, Q., Zhao, J., Dong, Y., Zhu, X., Lin, S., Zhang, H., Zhang, Z., 2021. Gap filling for historical Landsat NDVI time series by integrating climate data. Remote Sens. 13, 484. URL: http://dx.doi.org/10.3390/rs13030484, doi:10.3390/rs13030484.

Zemicheal, T., Dietterich, T.G., 2019. Anomaly detection in the presence of missing values for weather data quality control, in: Proc. ACM COMPASS, Accra, Ghana. p. 65–73. doi:10.1145/3314344.3332490.

Zhang, S., Zhu, Y., You, Z., Wu, X., 2017. Fusion of superpixel, expectation maximization and phog for recognizing cucumber diseases. Comput. Electron. Agric. 140, 338–347. URL: https://www.sciencedirect.com/science/article/pii/S0168169917302910, doi:https://doi.org/10.1016/j.compag.2017.06.016.