Crop monitoring and detection of anomalous crop development at the parcel-level with multispectral and synthetic aperture radar satellite data

Florian Mouret

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Informatique et Télécommunication

Présentée et soutenue par :

M. FLORIAN MOURET

le vendredi 4 février 2022

Titre :

Crop monitoring and detection of anomalous crop development at the parcel-level with multispectral and synthetic aperture radar satellite data

---

Ecole doctorale :

Mathématiques, Informatique, Télécommunications de Toulouse (MITT)

Unité de recherche :

Institut de Recherche en Informatique de Toulouse (IRIT)

Directeur(s) de Thèse :

M. JEAN-YVES TOURNERET

M. DENIS KOUAMÉ

Rapporteurs :

M. JEAN-PHILIPPE OVARLEZ, ONERA - CENTRE DE PALAISEAU

MME FLORENCE TUPIN, TELECOM PARISTECH

Membre(s) du jury :

M. FRÉDÉRIC PASCAL, CENTRALESUPELEC GIF SUR YVETTE, Président

M. DENIS KOUAMÉ, UNIVERSITE PAUL SABATIER, Membre

M. JEAN-YVES TOURNERET, TOULOUSE INP, Membre

MME SYLVIE DUTHOIT, TerraNIS, Membre

M. MOHANAD ALBUGHDADI, TerraNIS, Membre

M. PAULO GONCALVES, ENS SCIENCES LYON, Membre

Résumé —

La surveillance des cultures va devenir un enjeu majeur dans les années à venir. Soumise aux pressions liées au changement climatique d'une part, et à l'augmentation de la population mondiale d'autre part, les chaînes d'approvisionnement alimentaire risquent d'être fortement contraintes, impactant la sécurité alimentaire dans de nombreuses zones de la planète. Dans ce contexte, l'utilisation de la télédétection pour acquérir des informations sur l'état de la végétation sera un outil primordial. Un des domaines directement concerné est l'agriculture de précision, qui consiste à optimiser les rendements et les pratiques agricoles. Avec l'arrivée des satellites de la mission Copernicus, Sentinel-1 (radar à synthèse d'ouverture) et Sentinel-2 (imagerie multispectrale), les possibilités d'applications dans le domaine ont été décuplées. En effet, les données Sentinel sont disponibles gratuitement, et ce, avec une résolution temporelle et spatiale adaptée à la surveillance des cultures au niveau de la parcelle. L'objectif principal de cette thèse est de proposer une stratégie pour détecter automatiquement les parcelles agricoles qui ont un développement agronomique anormal. Une attention particulière a été donnée à l'utilisation conjointe des données Sentinel-1 et Sentinel-2. De plus, afin d'être déployée facilement dans un contexte opérationnel, une contrainte est de proposer une méthode capable d'analyser un seul cycle de croissance (ou une partie de celui-ci).

Pour répondre aux objectifs de la thèse, nous proposons dans un premier temps une chaîne de traitement permettant l'extraction d'indicateurs agronomiques au niveau de la parcelle. Ces indicateurs sont calculés en deux temps : 1) calcul d'indicateurs agronomiques au niveau pixel et 2) calcul de statistiques spatiales au niveau parcelle. Par la suite, ces indicateurs sont utilisés pour détecter des parcelles qui ont un comportement phénologique anormal. La détection est non supervisée et réalisée à l'aide d'un algorithme de détection d'anomalie. Une comparaison de plusieurs approches a été faite pour trouver la méthode la plus adaptée à notre problème. Parmi les différents algorithmes testés, la méthode la plus efficace est la forêt d'isolement, qui présente également l'avantage d'être rapide et peu sensible aux choix de ses paramètres. Grâce à la méthode proposée, il est possible de détecter des parcelles au comportement anormal avec une grande précision. Les résultats obtenus ont été validés sur deux types de cultures différentes, le blé et le colza. Dans un second temps, nous traitons le problème de détection d'anomalie en présence de données manquantes. Cette problématique est fondamentale en télédétection, en particulier pour les données issues d'images multispectrales car celles-ci sont sensibles au couvert nuageux. Pour résoudre ce problème, nous proposons de reconstruire les données manquantes (au niveau parcelle) en utilisant des modèles de mélange gaussien. Cette approche s'est montrée significativement meilleure que les autres approches testées pour reconstruire les données manquantes et pour permettre de détecter des anomalies sur des parcelles agricoles avec des séries temporelles incomplètes. De plus, nous avons également proposé une méthode d'estimation des modèles de mélange gaussien qui est robuste à la présence de valeurs aberrantes dans les données. Cette méthode est particulièrement utile en présence de forte valeurs anormales, par exemple en présence de parcelles provenant d'un type de culture différent de celui analysé. Enfin, nous

explorons dans cette thèse des approches de détection d'anomalie qui prennent en compte la structure temporelle des données. En particulier, nous proposons une méthode basée sur un ensemble de modèles de Markov cachés. Un des intérêts de cette approche est de pouvoir également localiser les anomalies dans le temps.

Mots clés : télédétection, imagerie multispectrale, imagerie radar, surveillance des cultures, détection d'anomalies, reconstruction de données manquantes (imputation).

Abstract —

Crop monitoring will become a major challenge in the coming years. Under the pressure of climate change on the one hand, and the increase of the world population on the other hand, food supply chains are likely to be strongly constrained, impacting food security in many areas of the planet. In this context, using remote sensing to acquire information on vegetation status will be a key asset. One of the areas directly concerned is precision agriculture, which consists in optimizing yields and agricultural practices. With the arrival of the Copernicus mission satellites, Sentinel-1 (synthetic aperture radar) and Sentinel-2 (multispectral imagery), the possibilities of applications in this area have increased drastically. Indeed, Sentinel data are freely available, with a temporal and spatial resolution adapted to crop monitoring at the parcel level. The main objective of this thesis is to propose a strategy to automatically detect agricultural parcels with abnormal agronomic development. Special attention was given to the joint use of Sentinel-1 and Sentinel-2 data. Moreover, in order to be easily deployed in an operational context, a constraint is to have a method able to analyzing a single growth cycle (or a part of it).

To meet the objectives of the thesis, we first propose a processing chain allowing the extraction of agronomic indicators at the parcel-level. These indicators are calculated in two steps: 1) calculation of agronomic indicators at the pixel level and 2) calculation of spatial statistics at the plot level. Then, these indicators are used to detect parcels with abnormal phenological behavior. The detection is unsupervised and performed using an anomaly detection algorithm. A comparison of several approaches was made to find the most suitable method for our problem. Among the different algorithms tested, the most efficient method is the isolation forest, which also has the advantage of being fast and not very sensitive to the choice of its parameters. Thanks to the proposed method, it is possible to detect plots with abnormal behavior with a high accuracy. The results obtained were validated on two different types of crops, wheat and rapeseed. In a second step, we addressed the problem of anomaly detection in the presence of missing data. This problem is fundamental in remote sensing, in particular for multispectral data because they are sensitive to cloud cover. To solve this problem, we propose to reconstruct the missing data (at the parcel-level) using Gaussian mixture models. This approach has been found to be significantly better than the other tested approaches for reconstructing missing data and for detecting anomalies on parcels with incomplete time series. In addition, we also have proposed a method for estimating Gaussian mixture models that are robust to the presence of outliers in the data. This method is particularly useful in the presence of strong outlier values, for example in the presence of parcels coming from a different crop type than the one analyzed. Finally, we explore in this thesis anomaly detection approaches that take into account the temporal structure of the data. In particular, we propose a method based on an ensemble of hidden Markov models. One of the interests of this approach is to be able to localize the anomalies in time.

Keywords: remote sensing, multispectral imagery, radar imagery, crop monitoring, anomaly detection, data imputation.

Remerciements

Ce travail de thèse a été rendu possible grâce au soutien de nombreuses personnes que j'aimerais remercier ici. Je souhaite tout d'abord exprimer ma gratitude envers mes encadrants de thèse, tant du côté académique que du côté de l'entreprise TerraNIS. Je tiens particulièrement à remercier mon directeur de thèse, Jean-Yves Tourneret, qui a été d'un soutien sans faille durant ces trois années de doctorat et a su se rendre disponible pour répondre à mes (nombreuses) questions. Jean-Yves, tu m'as permis d'apprendre une méthode de travail rigoureuse et précise, tout en me motivant à aller au bout des choses. J'aimerais enfin te remercier pour les qualités humaines dont tu as fait preuve tout au long de la thèse, un point qu'il me semble important de mettre en avant car il a été structurant durant tout mon doctorat. Merci également à Denis Kouamé, Sylvie Duthoit et Mohanad Albugdhadi pour leur encadrement et conseils. Vous avez pu, chacun à votre manière, me guider tout au long de la thèse. Denis, merci d'avoir partagé tes connaissances théoriques et ton expérience académique afin d'améliorer mes travaux. Sylvie, merci d'avoir partagé ton savoir-faire et tes compétences thématiques qui m'ont permis d'aborder mes problématiques sous un autre angle. Je tiens aussi à te remercier pour ta gentillesse et ta disponibilité. Mohanad, merci d'avoir partagé tes connaissances techniques et académiques. J'ai grandement apprécié les moments passés ensembles à TerraNIS (et en dehors).

Je remercie chaleureusement toute l'équipe de TerraNIS pour son accueil. J'ai apprécié travailler avec vous, toujours dans une ambiance chaleureuse et agréable. Cet environnement de travail a été un point positif lors de ces trois années. J'aimerais particulièrement remercier David et Marc d'avoir permis cette aventure en m'engageant comme doctorant. Merci à Guillaume d'avoir pris le temps de me fournir une expertise agronomique de qualité, ainsi que pour les bons moments passés à parler d'autres choses que de boulot. Je remercie Ève, avec qui j'ai partagé cette expérience de doctorat à TerraNIS, pour ses conseils et sa bonne humeur. Enfin, merci à Maïlys, aux Nicos, Camille, Quentin, Cécile, Anne, Audrey, Clément, Nathalie, Corinne, Thomas... pour les discussions passionnantes et les bons moments passés ensemble.

Je remercie toute l'équipe TéSA de m'avoir accueillie avec bienveillance lorsque je venais voir Jean-Yves. En particulier, merci Corinne pour ta gentillesse et ta bonne humeur. Je remercie enfin Adrien, Barbara, Kareth, Oumaima, Julien, Selma, Corentin, Victor, Isabelle, Philippe, ... pour votre accueil qui m'a permis de passer de bons moments.

Je remercie mes rapporteurs de thèse, Jean-Philippe Ovarlez et Florence Tupin, pour le temps qu'ils ont consacré à la relecture de ce manuscrit ainsi que pour leurs remarques et conseils avisés. Je remercie également Paulo Gonçalves et Fréderic Pascal d'avoir accepté d'être membres de mon jury de thèse. Je souhaiterais également remercier Frédéric, ainsi que Alexandre Hippert-Ferrer, pour le temps consacré lors de notre collaboration, que j'ai trouvé très enrichissante.

Je tenais aussi à remercier toutes les personnes du monde académique et industriel qui

ont pu me donner leurs avis et conseils au cours des échanges que j'ai pu avoir avec eux. Je pense en particulier à Hervé Poilvé, Milena Planells, Stéphane Mermoz, Alexandre Bouvet, Axel Carlier, Esa Ollila, Nabil El Korso, ... j'en oublie sûrement.

Je remercie ma famille pour son soutien sans faille, et en particulier mes parents qui m'ont permis d'arriver jusque là. Je remercie aussi mes amis, à Toulouse et ailleurs, de m'avoir permis d'oublier quelques instants les tracas du quotidien. Dr Coutet/Poutine, maintenant tu n'es plus le seul à pouvoir avoir une plaque sur ton bureau !

Mes derniers remerciements sont évidemment pour Manon. Si j'ai fait tout cela, c'est grâce à toi.

Contents

List of Figures

List of Tables

Acronyms

AE	*Autoencoder*
AUC	*Area Under the Curve*
EM	*Expectation-Maximization*
ESA	*European Space Agency*
IF	*Isolation Forest*
IQR	*Interquartile Range*
fCover	*Fractional vegetation cover*
FP	*False Positive*
GMM	*Gaussian Mixture Model*
GRD	*Ground Range Detected*
GRVI	*Green Red Vegetation Index*
HMM	*Hidden Markov Model*
KNN	*K-Nearest Neighbors*
LAI	*Leaf Area Index*
LOF	*Local Outlier Factor*
LoOP	*Local Outlier Probabilities*
LPIS	*Land Parcel Identification System*
MAE	*Mean Absolute Error*
MAJA	*MACCS ATCOR Joint Algorithm*
MC	*Monte Carlo*
MCARI	*Modified Chlorophyll Absorption Ratio Index*
MICE	*Multiple Imputation by Chained Equation*
NDVI	*Normalized Difference Vegetation Index*
NDWI	*Normalized Difference Water Index*
NIR	*Near Infra-Red*

OCSVM	*One Class SVM*
OSAVI	*Optimized Soil Adjusted Vegetation Index*
Radar	*RAdio Detection And Ranging*
RVI	*Radar Vegetation Index*
S1	*Sentinel-1*
S2	*Sentinel-2*
SAR	*Synthethic Aperture Radar*
SHAP	*SHapley Additive exPlanations*
SVM	*Support Vector Machine*
TICC	*Toeplitz Inverse Covariance-Based Clustering*
TP	*True Positive*
VI	*Vegetation Indice*

Chapter 1

Introduction

Contents

1.1 Remote sensing for agriculture

1.1.1 General context

Farming is expected to feed more than 10 billions people by 2050, increasing agricultural demand by 50% compared to 2013 under modest growth scenario (FAO, 2017). This boost in the food production is challenged by climate change, which will affect food security at various stages of the food chain (Tirado et al., 2010; Wheeler and von Braun, 2013). In addition, a change in agricultural practices is needed to decrease their negative impact on biodiversity (Newbold et al., 2015), water resources and green-house gas emissions (Gomiero et al., 2011).

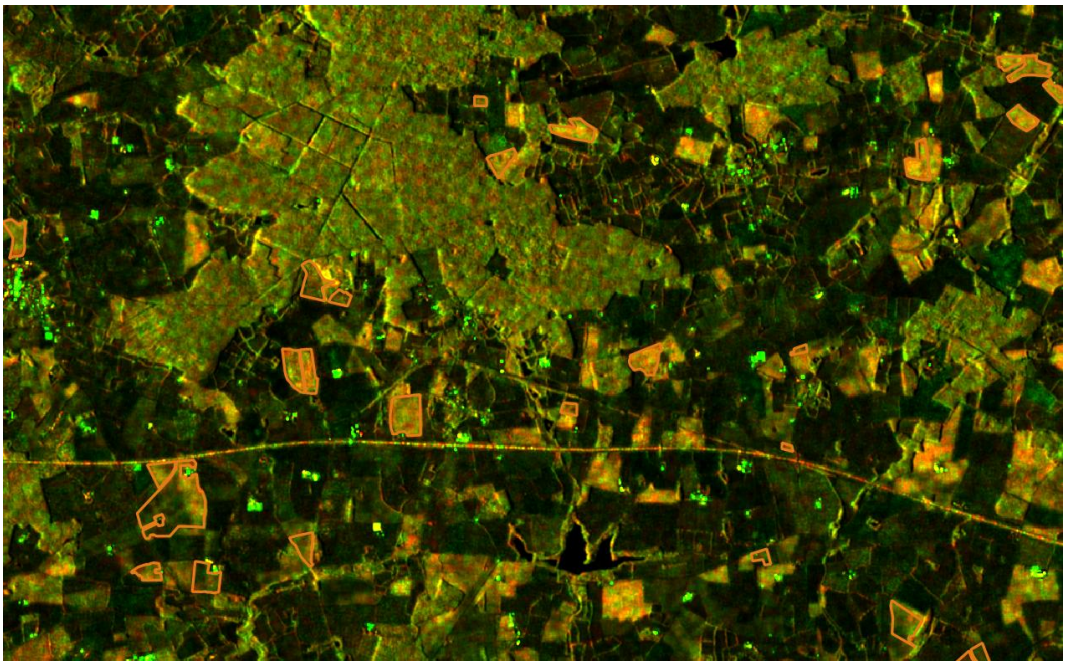
In this context, monitoring crop growth and status becomes a necessary issue for a wide variety stakeholders (Weiss et al., 2020). Remote sensing can provide critical information to the agricultural sector in a timely and reliable manner at large-scale (Atzberger, 2013), which is important to measure sustainable intensification and optimize crop practices (Areal et al., 2018). Moreover, near real-time monitoring can help to improve food system resilience and react to extreme events (Wheeler and von Braun, 2013).

The different remote sensing applications in agriculture can be grouped into 4 categories: phenotyping, yield forecasting, ecosystem services and precision farming (Weiss et al., 2020). Precision agriculture, which is the focus of this thesis, aims at monitoring crops for the optimization of yields as well as farming practices. It covers a wide range of applications, such as weed and disease detection (López-Granados, 2011; Mahlein, 2016) and nutrient and water stress monitoring (Baret et al., 2007; Calera et al., 2017). The use of remotely sensed imagery for precision agriculture is particularly interesting because it provides spatial and temporal information about the condition of crops in a non-destructive manner and without needing on-site visits (Schulz et al., 2021).

1.1.2 Precision agriculture with Sentinel-1 and Sentinel-2 satellites

Until recently, the current limitations for image-based remote sensing applications were mainly due to sensor attributes (*e.g.*, spectral range), spatial resolution and revisit time (Moran et al., 1997). Nowadays, the amount of freely accessible remote sensed images has drastically increased, thanks to the Copernicus mission of the European Union operated by the European Space Agency (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, respectively (Torres et al., 2012). Sentinel-1 (S1) and Sentinel-2 (S2) satellites have a high temporal and a spatial resolution. This is adapted to work at the parcel level (for very high resolution analysis, *e.g.*, at the plant-level, these resolutions are however not sufficient). Both types of sensors are complementary and have been largely studied for this application. An example illustrating the information available using S1 and S2 images is displayed in Figure 1.1.

(a)



(b)

Figure 1.1: (a) Sentinel 2 image acquired on June 28, 2018 (true colors) and (b) Sentinel-1 image acquired on June 25, 2018 (multitemporal speckle filtering was applied, composite colors: Red=VH, Green=VV). Orange boundaries are rapeseed crop fields to be monitored.

1.1.2.1 Sentinel-2: multispectral imaging satellites

S2-A and S2-B are multispectral imaging satellites with 13 spectral bands covering the visible, the near infra-red (NIR) and the shortwave-infrared (SWIR) spectral regions (Drusch et al., 2012). Details about the S2 spectral bands are reported in Table 1.1, where it can be seen that the different spectral bands have different spatial resolutions (from 10m to 60m). When using both S2-A and S2-B satellites, a theoretical revisit time of 5 days can be reached.

Table 1.1: Sentinel-2 multispectral bands. NIR refers to Near Infrared whereas SWIR refers to Shortwave Infrared.

Spectral bands	Central wavelength (μm)	Bandwidth (μm)	Resolution (m)
Band 1: Coastal aerosol	0.443	0.021	60
Band 2: Blue	0.492	0.066	10
Band 3: Green	0.560	0.035	10
Band 4: Red	0.665	0.030	10
Band 5: Vegetation Red Edge	0.705	0.015	20
Band 6: Vegetation Red Edge	0.740	0.015	20
Band 7: Vegetation Red Edge	0.783	0.020	20
Band 8: NIR	0.842	0.115	10
Band 8: Narrow NIR	0.864	0.021	20
Band 9: Water vapour	0.945	0.020	60
Band 10: SWIR - Cirrus	1.374	0.031	60
Band 11: SWIR	1.610	0.090	20
Band 12: SWIR	2.202	0.175	20

Multispectral images have been used for decades thanks to their convenient interpretation and exploitation. Many Vegetation Indices (VI) have been proposed to provide a simple and easy evaluation of the vegetation cover (Bannari et al., 1995). One can mention the famous Normalized Difference Vegetation Index (NDVI) (Rouse et al., 1974), which is still widely used nowadays to monitor vegetation health and vigor (Meroni et al., 2019; Garioud et al., 2021). Bio-physical properties of the vegetation, *e.g.*, the Leaf Area Index (LAI) or the fraction of green vegetation cover (fCover) (Djamai et al., 2019; Verrelst et al., 2015), can also be retrieved with multispectral images. Finally, machine learning methods can take advantage of the various spectral bands to perform tasks such as land cover classification (Gómez et al., 2016; Inglada et al., 2017). In this context, the use of S2 satellites is a great opportunity to use multispectral data in a consistent manner (Segarra et al., 2020).

1.1.2.2 Sentinel-1: SAR C-band imaging satellites

S1-A and S1-B are SAR C-band imaging satellites whose center frequency is 5.405 GHz. When using both S1-A and S1-B satellites, a theoretical revisit time of 6 days can be reached. S1 images are available in dual polarization (VH+VV) with a 5×20 m spatial resolution (in Interferometric Wide (IW) swath mode). SAR remote sensing is sensitive to the dielectric and geometrical characteristics of the plant, as illustrated in Figure 1.2. SAR data are available for any sunlight and cloud coverage conditions, unlike optical images that are sensitive to these phenomenona (Wang et al., 2009).

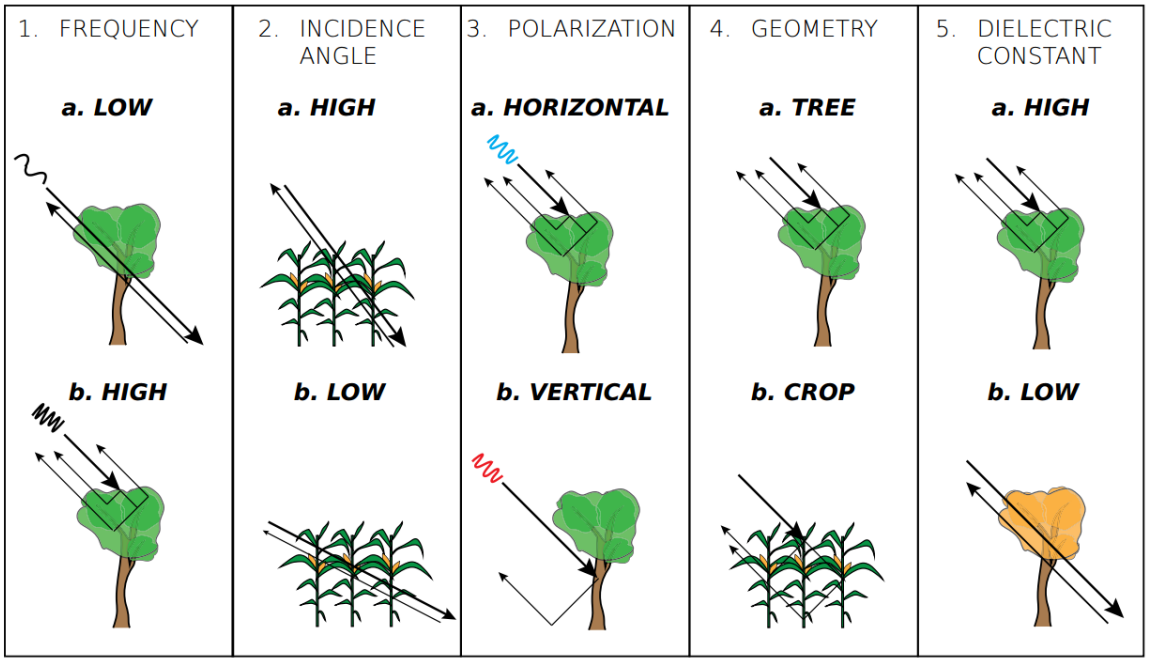


Figure 1.2: Illustrative examples of different factors influencing the SAR backscatter. Arrows indicate the radar waves and line width represents higher or lower backscattering (van Emmerik, 2017).

SAR data have been used for instance for crop monitoring (Betbeder et al., 2016; Khabbazan et al., 2019), crop mapping (Abdikan et al., 2016) or water stress detection (van Emmerik, 2017). As an example, SAR data have proven to be well suited to rice monitoring, particularly because rice crops grow in cloudy or foggy areas (Bouvet et al., 2009; He et al., 2018). A Detailed review on the different applications of SAR in agriculture is available in (Liu et al., 2019), which concludes on the great potential of SAR data in the various fields of agricultural remote sensing.

1.1.2.3 On the complementarity of S1 and S2 data

The complementarity of SAR and multispectral images has been used to address problems including crop type classification (Inglada et al., 2016; Denize et al., 2018; Kussul et al., 2018; Campos-Taberner et al., 2019; Orynbaikyzy et al., 2019), estimation of crop water requirement (Navarro et al., 2016) and change detection (Prendes et al., 2015a,b,c). In most of these studies, using additional SAR data was found important to provide complementary information on the vegetation status. For instance, Campos-Taberner et al. (2019) observed better classification scores when using additional S1 data, especially for the most confusing classes. Inglada et al. (2016) pointed out the interest of using SAR data to complete sparse mutlispectral time series (due to cloud coverage). A comprehensive analysis of the temporal behavior of S1 and S2 data has also been proposed (Veloso et al., 2017), concluding on the unique opportunity to monitor crops systematically using these data. Finally, in a recent

study Meroni et al. (2021) demonstrated that both S1 and S2 data can provide relevant and at times complementary land surface phenology (LSP) information at field and crop-level.

All these factors (free access, adapted temporal/spatial resolutions and good properties for crop monitoring) motivated to use S1 and S2 sensors for this thesis.

1.2 Detecting anomalies in the vegetation status: state-of-the-art

One remaining challenge in precision agriculture is the automatic detection of agricultural parcels that have an abnormal vegetation development. An illustrative example is displayed in Figure 1.3, where several crop fields are visualized using S2 images acquired on (a) February 25, 2018 and (b) April 21, 2018. At this stage of the growing season (end of winter / beginning of the flowering stage), one can notice that some crop fields (here rapeseed crops) are more or less affected by heterogeneity and that this heterogeneity can be more or less transient. Detecting parcels whose phenological behaviors significantly differ from the others could help users such as farmers or agricultural cooperatives to optimize agricultural practices, disease detection or fertilization management. It could also be valuable in areas such as subsidy control or crop insurance.

(a) (b)

Figure 1.3: Rapeseed parcels (red and yellow boundaries) visualized using S2 images acquired on (a) February 25, 2018 and (b) April 21, 2018 (true colors).

This section first provides a general state of the art on outlier detection, mainly to present the main concepts and challenges related to this area. In a second step, a focus is made on the literature related to the detection of anomalies in the vegetation status. Finally, based on this literature, we motivate a need to find new methods adapted to the specific case of crop monitoring at the parcel-level.

1.2.1 Outlier detection: general survey

In the literature, the problem of finding samples that are unusual or different from the majority of the data is known as outlier detection (also referred to as anomaly detection). Outlier detection techniques have received a considerable attention (Aggarwal, 2017; Chandola et al., 2009) since they are used in a large variety of application domains, *e.g.,* fraud detection or medical diagnosis. Outlier detection algorithms typically provide an *outlier score* for each sample of the dataset, which quantifies the degree of abnormality of the samples (some algorithms only provide binary labels). Outlier scores can be converted to binary labels for the final decision making, for instance by detecting a percentage of the samples with the highest scores. This percentage is known as the *outlier ratio*.

Outliers or anomalies are samples that do not conform to expected behavior. Outlier detection seems to be a simple task, but several factors make this approach very difficult in practice (Chandola et al., 2009). First, the boundary between normal data (also called inliers) and anomalies is generally subjective or not precise, as illustrated in Figure 1.4. Moreover, anomalies are specific to each application domain and labeled data are generally not available. Therefore, in practice the outlier detection strategy must be designed specifically for the problem at hand, generally in a fully unsupervised mode.

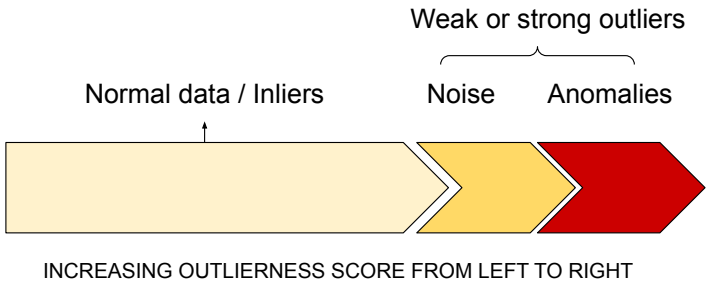


Figure 1.4: The spectrum from normal data to outliers (Aggarwal, 2017).

Three types of anomalies have been defined by Chandola et al. (2009):

- *Point anomalies:* a sample considered anomalous with respect to the rest of the data is a point anomaly. Hence, point anomaly detection algorithms generally aim at comparing each instance to the rest of the data to find the most isolated / far from the others samples. Most of the literature on outlier detection focuses on finding point anomalies.
- *Contextual anomalies:* if a value is considered anomalous in a specific context (but not otherwise), it is a contextual anomaly. To find such anomaly, contextual attributes (e.g., typically spatial or temporal information) are mandatory since they are needed to separate outlier values from the normal ones.
- *Collective anomalies:* if a set of samples is anomalous with respect to the entire dataset whereas each individual sample of this set is not anomalous by itself, it is considered as a collective anomaly. A typical example is a sequence of actions occurring in a

computer being victim of a remote attack. Generally, each element of the sequence is not anomalous by itself whereas the complete sequence should be detected as a threat.

Anomalies occurring in remote sensing time series are either contextual (i.e., a unique time stamp is anomalous) or collective (large sub-sequence or the entire time series is anomalous) anomalies (Aggarwal, 2017, Chapter 9). In all cases, the time attribute is crucial to determine which values of the time series are anomalous. An example illustrating the importance of time information is displayed in Figure 1.5, which shows the median NDVI time series of a rapeseed parcel (orange line) which is compared to the whole dataset (the blue area is filled between the 10th and 90th percentiles of all the other analyzed rapeseed parcels). One can observe that having a median NDVI close to 0.5 is not anomalous by itself, but can be anomalous at specific times of the growing season. Moreover, while a unique abnormal value (contextual anomaly) could be related to noise or acquisition problem (*e.g.*, undetected cloud), having most of the time series abnormal (collective anomaly) can indicate a strong problem in the parcel behavior.

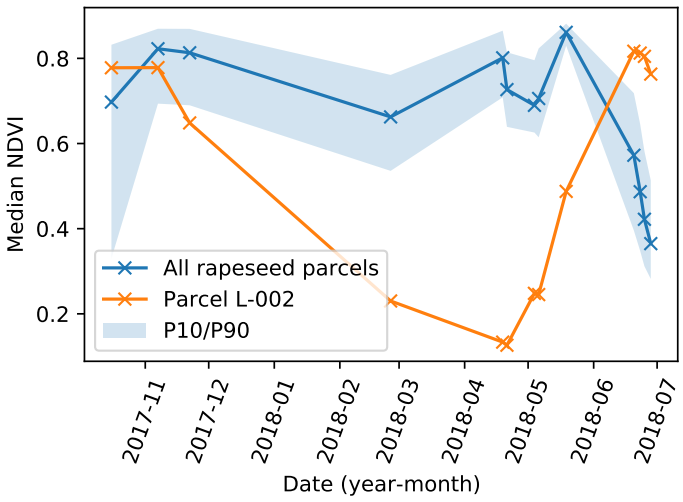


Figure 1.5: Median NDVI time series of an outlier rapeseed parcel (orange line). The blue line is the median value of the whole dataset composed of 2218 rapeseed parcels. The blue area is filled between the 10th and 90th percentiles.

Two main approaches can be used to detect anomalies in multidimensional time series, e.g., formed by the concatenation of several vegetation indices such as the NDVI:

- *Prediction-based techniques* (Aggarwal, 2017, Chapter 9.2): these techniques aim at modeling the normal temporal behavior of the data, and use this dynamic model to detect anomalies. If an observed value does not correspond to the value predicted by the model, it is detected as an anomaly. Such technique is well suited to detect contextual anomalies, which are disturbances deviating from the expected forecasting.
- *Time series of unusual shapes and reduction to the point anomaly detection problem* (Aggarwal, 2017, Chapter 9.3): such approach is adapted to detect collective anomalies

that lead to have time series with unusual *shapes* when compared to the underlying series (it is for instance the case in the example provided in Figure 1.5). This problem is generally addressed by a reduction to the point anomaly detection problem, *e.g.*, by considering each time series as a feature vector to be compared with the rest of the data. By doing this, one can take advantage of the various point anomaly detection algorithms available in the literature.

In the next part of this section, we will see that the detection of anomalies in remote sensing time series have been mostly addressed by using prediction-based techniques. Then, we will explain why these techniques are not particularly well suited to our use case and motivate the need for a new detection strategy.

1.2.2 Application to remote sensing data for the analysis of the vegetation status

In Earth observation, the majority of the studies have been devoted to the detection of abnormal vegetation status at the country-level using time series constructed from the NDVI. Most of these approaches are *prediction-based* techniques (as defined in the previous section), and aim at modeling NDVI time series using historical data and detecting potential anomalies by comparing new observations with their corresponding predicted values.

Common approaches are based on a parametric model for the time evolution of NDVI such as the double logistic or symmetric Gaussian models (Atzberger and Eilers, 2011a; Beck et al., 2006). Smoothing techniques can help to have more reliable time series and to work in near real time (NRT) (Atzberger and Eilers, 2011b; Hird and McDermid, 2009; Klisch and Atzberger, 2016; Meroni et al., 2019). Various other approaches have been investigated using season trend models (Verbesselt et al., 2012; Zhou and Tang, 2016), Seasonal Autoregressive Integrated Moving Average (SARIMA) models (Zhou et al., 2016) and prediction models such as extended Kalman filters (Sedano et al., 2015). Finally, recent studies have investigated similar techniques for S2 data, as for instance in Kanjir et al. (2018) where Breaks for Additive Season and Trend (BFAST) are used to detect land use anomalies. Note that the BFAST technique was introduced earlier by Verbesselt et al. (2010) to monitor phenological change detection in NDVI time series.

1.2.3 A need for new strategies adapted to precision agriculture and parcel-level analysis

The aforementioned approaches can be difficult to implement for our specific use case, which consists of detecting abnormal development in parcels of a given crop type. First, modeling the normal behavior of the data implies having access to normal representative examples, which can be challenging and time consuming in practice. Crop rotation, lack of historical data and the inconsistency of S2 time series due to the cloud coverage are other factors leading to a harder practical implementation. Moreover, forecasting techniques need long historical data

to be fitted efficiently. In our case, analyzing a unique growing season is more appropriate, mainly to facilitate operational service and costs for practical applications (having access to reliable parcel data coming from multiple growing seasons to build accurate temporal models is another major issue). Finally, while most of the studies focus only on the analysis of the NDVI, it seems relevant to use a larger variety of indicators coming from S1 and S2 data. This literature overview motivates the need to investigate new approaches for outlier detection dedicated to crop monitoring.

1.3 Problem formulation and objectives of the thesis

This thesis aims to explore the challenge of automatic crop monitoring using S1 and S2 data. More precisely, it aims at detecting agricultural parcels whose phenological behavior significantly differs from the others. Since the study is conducted at the parcel-level for a specific crop type, one hypothesis is that parcel boundaries and the crop type are available a priori. The temporal analysis was limited to a single growing season, mainly to have an easier operational deployment (as explained in the previous section), to avoid theoretical issues (caused by crop rotation, time series inconsistencies, etc). For the same reasons, the proposed method has to be the most unsupervised possible (*e.g.*, regarding parameter tuning and need for labeled data, etc) and robust to changes (*e.g.*, in crop types, data types, etc). Another important challenge, which is recurrent in remote sensing, is the need to deal with missing data (coming from clouds for S2 images or acquisition problems). Finally, considerations related to the end-user needs have to be taken into account (*e.g.*, interpretability of the detection results, adaptation to the user feedback, etc.). The main objectives and challenges of the thesis are summarized in what follows:

→ Detect relevant anomalies (*i.e.*, related to agronomic phenomenon) at the parcel-level.

→ Give a relevant score of abnormality for each parcel (*i.e.*, strong anomalies have a higher outlier score).

→ Analyze the crop parcels within a single growing season (or a part of the growing season if possible).

→ Use efficiently the complementary of S1 and S2 data.

→ Validate the method on various crop types.

→ Propose a fully automated method (no need for manual labeling and parameter tuning).

→ Handle missing data (*e.g.*, due to images partially covered by clouds or acquisition problem).

1.4 Organization of the manuscript

The main steps of the proposed approach are summarized in the diagram presented in Figure 1.6 and described in the first chapters of this manuscript. The remaining of this thesis is organized as follows:

- Chapter 2 is devoted to the data preprocessing and the feature extraction. Since the proposed method is fully unsupervised, having relevant features is needed. Moreover, for post-analysis, a particular attention has to be given to their interpretability. The main contribution of this chapter is to propose a feature extraction procedure that provides parcel-level features which characterize efficiently the parcel behavior, in terms of vigor and heterogeneity.
- Chapter 3 introduces a strategy to detect agricultural parcels with anomalous phenological development. Various outlier detection methods are compared and the effects of changes in the features set and configuration are analyzed. The main contribution of this chapter is to propose a fully unsupervised method for the detection of parcels with abnormal development. Another important contribution is the systematic characterization of the different anomalies observed, which have been grouped in different categories. The main results of this chapter have been published in Mouret et al. (2021a).
- Chapter 4 focuses on the reconstruction of missing data, which is a recurrent problem in remote sensing. The main contribution of this chapter is to propose a Gaussian Mixture Model (GMM) imputation strategy, which is very competitive with respect to the existing reconstruction methods. Moreover, a new robust GMM is proposed to take into account the presence of irrelevant samples in the dataset. The main results of this paper are currently under review for publication (Mouret et al., 2021b).
- Chapter 5 focuses on techniques based on the analysis of time series for the detection of anomalous crop development at the parcel-level. These experiments aim to challenge the method proposed in Chapter 3. Since they rely on a temporal analysis, these methods generally provide additional outputs that can be interesting for crop monitoring (temporal localization of the anomalies, temporal state of the analyzed parcels, etc).
- Chapter 6 provides a general conclusion to this manuscript.

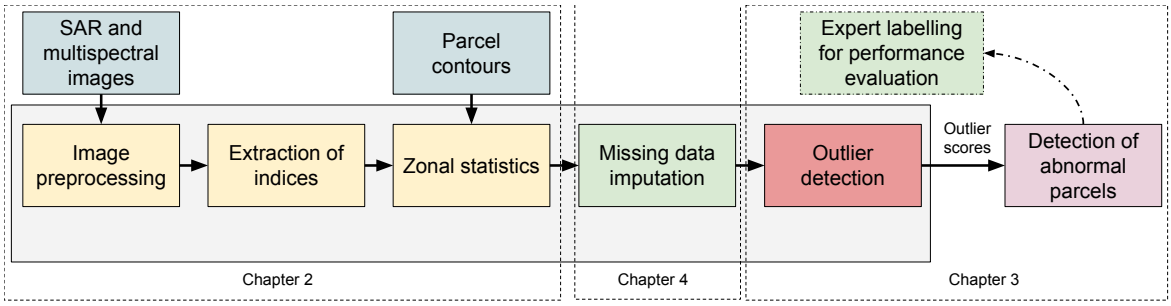


Figure 1.6: Diagram summarizing the methodological steps for the detection of anomalous crop development.

International journal papers:

1. **Mouret F**, Albughdadi M, Duthoit S, Kouamé D, Rieu G, Tourneret J-Y. Outlier detection at the parcel-level in wheat and rapeseed crops using multispectral and SAR time series. *Remote Sensing*. 2021; 13(5):956.

2. León-López K, **Mouret F**, Arguello H, Tourneret J-Y. Anomaly Detection and Classification in Multispectral Time Series based on Hidden Markov Models. *IEEE Transactions on Geoscience and Remote Sensing*. 2021.

3. **Mouret F**, Albughdadi M, Duthoit S, Kouamé D, Rieu G, Tourneret J-Y. Reconstruction of Sentinel-2 Derived Time Series Using Robust Gaussian Mixture Models —Application to the Detection of Anomalous Crop Development. *Under review*. 2021.

Chapter 2

Data processing for the extraction of parcel-level features

Contents

2.1 Introduction

This section presents the study area, parcel data and remote sensing data used in the different experiments conducted throughout the thesis. A particular attention is devoted to the processing of the remote sensing data. Indeed, using raw remote sensing data is generally not suitable for two main reasons: 1) data preprocessing can improve the quality of the remote sensed images (*e.g.*, correction of atmospheric effects for S2 images (Hagolle et al., 2015), calibration and terrain correction for S1 images), 2) having features whose interpretation is easy can help to understand the crop behavior in a straightforward way. Moreover, extracting relevant features from these data is better suited to work with machine learning algorithms, especially in the unsupervised case.

2.2 Study area and parcel data

2.2.1 Study area

The analyzed area is located in the Beauce region in France. The area has an extent of 109.8×109.8 km^2 and is centered approximately at 48°24′N latitude and 1°00′E longitude (corresponding to the T31UCP S2 tile). Figure 2.1 shows the tile location and the studied area, which was chosen due to its richness of large crop fields such as wheat and rapeseed.

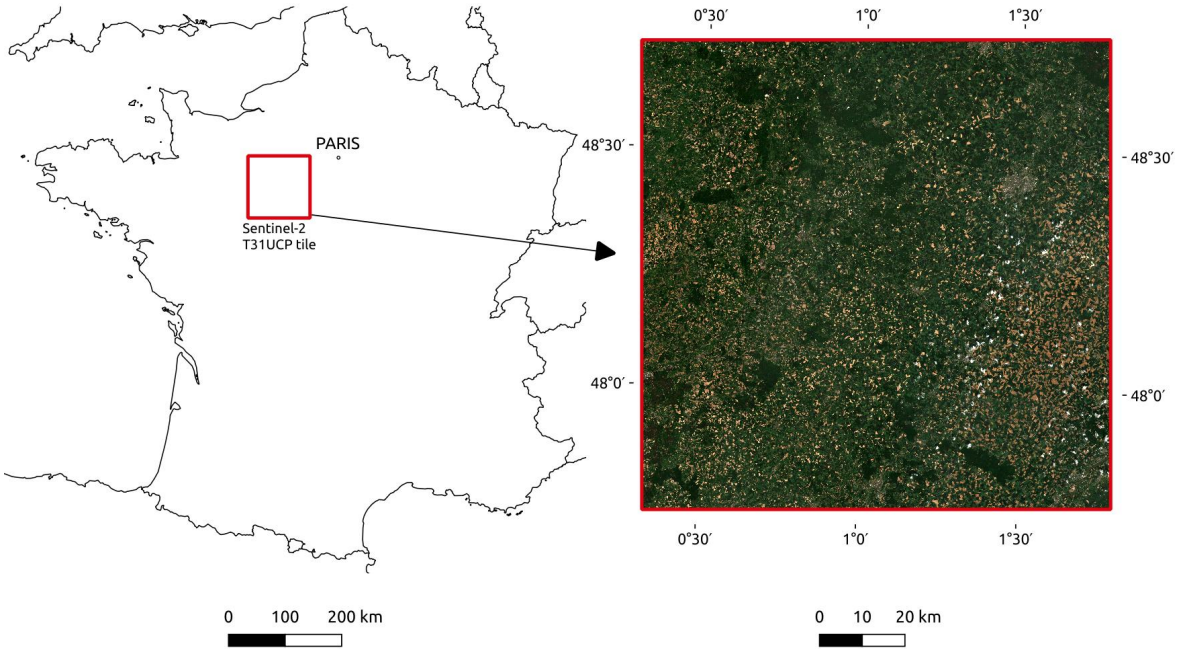


Figure 2.1: The Sentinel-2 tile T31UCP considered in this work is located in the Beauce area (near Paris) and delimited by the red box. On the right, the S2 image processed in level 2A acquired on May 19 2018 is displayed in natural colors.

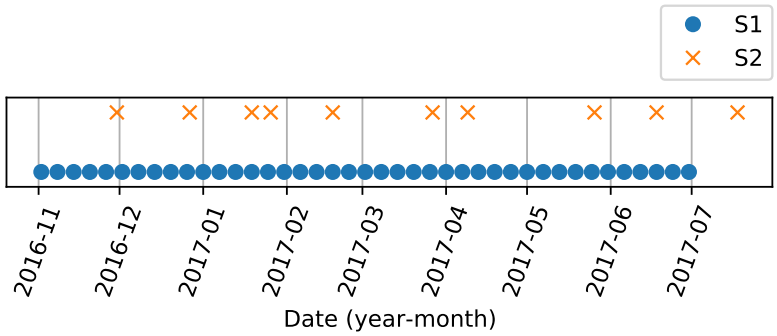
2.2.2 Parcel boundaries

The analysis is conducted on a total of 2218 rapeseed parcels (associated with the 2017/2018 growing season) and 3361 wheat parcels (associated with the 2016/2017 growing season). Parcel delineations are defined using a customer database, which allows us focusing on crop anomalies rather than detecting anomalies related to information coming from the database (such as the reported crop type, reported field delineation, etc.). To avoid problems in parcel boundaries, a buffer of 10 m was applied allowing too small parcels (area less than 0.5 ha) to be discarded from the database. All the parcels affected by clouds were discarded in a first analysis in order to have a reliable ground-truth. The French Land Parcel Identification System (LPIS) (Barbottin et al., 2018), which is available in open license, was also used to validate the robustness of our analysis to changes in parcel boundaries. This database is however generally available with a delay of 2 years, which is problematic for an operational service. An example of parcel delineations is provided in Figure 2.2 (another example is also provided in Figure B.8).

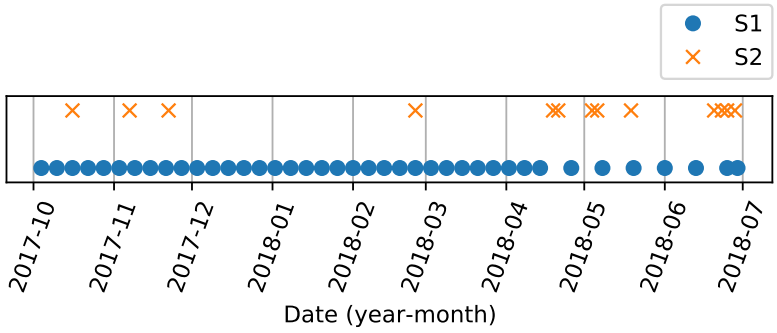
Figure 2.2: Example of parcel boundaries (rapeseed crop, growing season 2017/2018). In orange: customer database, in green: LPIS database.

2.3 Remote sensing data

The acquisition dates of S1 and S2 images are depicted in Figure 2.3 for the 2016/2017 and 2017/2018 growing seasons. It was decided to select S2 images with a low cloud coverage (cloud coverage lower than 20%). The strategy considered to handle remaining clouds is detailed in the next section. Regarding the S1 images, we used Ground Range Detected (GRD) products in the Interferometric Wide (IW) swath mode: phase information are lost compared to Single Look Complex (SLC) products but the volume of data is considerably reduced (which is a huge advantage for an operational service). For the 2016/2017 growing season, 41 S1 images and 10 S2 images were selected whereas 40 S1 images and 13 S2 images were selected in 2017/2018. Due to cloud coverage, the acquisition dates for S2 images are very different for the two growing seasons. Note that a reduced number of S1 images was available between May and July 2018. The absence of data during this period can be observed in all S1 data providers, which confirms problems in data acquisition.



(a)



(b)

Figure 2.3: Each marker corresponds to the acquisition date of a used image for the growing season (a) 2016/2017 and (b) 2017/2018.

2.3.1 Preprocessing of Sentinel-1 data

To build the database of S1 images, an offline processing inspired by the workflow proposed by Filipponi (2019) (illustrated in Figure 2.4) was conducted with the Sentinel Application Platform (SNAP, version 7.0)[1].

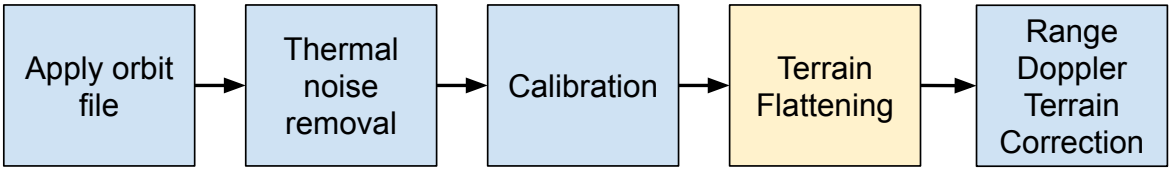


Figure 2.4: Sentinel-1 preprocessing chain used in the Sentinel Application Platform (SNAP). The yellow box, terrain flattening, was added to the workflow proposed in Filipponi (2019) to take into account the local incidence angle.

The different processing steps are detailed in the following:

- **Apply orbit file:** metadata of SAR products contain orbit state vectors that can be inaccurate. The precise orbit of the products can be updated using SNAP to address this issue.
- **Thermal noise removal:** S1 images can be corrupted by additive thermal noise (especially in areas with low backscatter).
- **Calibration:** this operation converts digital pixel values to radiometrically calibrated SAR backscatter. A calibration to β^0 backscatter (also called radar brightness) was used, since it is required to apply the terrain flattening operation. β^0 corresponds to the reflectivity per unit area in slant range.
- **Terrain flattening:** SAR backscattering is generally calibrated to σ^0 by considering the incidence angle θ (as proposed for instance in Filipponi (2019)). Instead, we used a terrain flattening operation that takes into account the local incidence angles (denoted as θ_l in Figure 2.5), since the analyzed area is wide and parcel features are compared to each other (the interest of using the local incidence is illustrated in Figure 2.5). This operation uses the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM) to produce γ^0 backscattering coefficients (Small, 2011).
- **Range Doppler terrain correction:** this operation provides orthorectified images to compensate for the distortions caused by the acquisition angle.

Note that a multi-temporal speckle filtering step was also tested without significant differences on the results (we implemented our own Python version of the filter introduced in Eq. (14) of Quegan and Jiong Jiong Yu (2001)). The best results were obtained with the workflow of Figure 1.6.

[1] `http://step.esa.int/main/toolboxes/snap/`, online accessed 8 December 2020

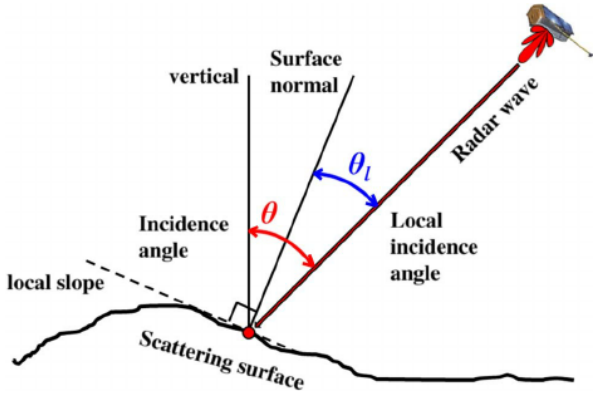


Figure 2.5: Reference geometry for the incidence angle θ and for the local incidence angle θ_l (Rizzoli and Bräutigam, 2014).

2.3.2 Sentinel-2 data preprocessing

S2 images were preprocessed using the MAJA processing chain (Hagolle et al., 2015) available on the PEPS platform of CNES. This preprocessing step provides level-2A ortho-rectified products expressed in surface reflectance. In addition to atmospheric correction, level-2A images are available with a cloud and shadow mask discarding irrelevant pixels in the images. A resampling strategy was adopted to obtain a spatial resolution of 10 m for the channels with a lower spatial resolution. Parcels fully covered by clouds during at least one time instant were discarded from the database and parcels partially covered by clouds were analyzed using pixels not covered by the cloud mask (the shadow mask was used in a similar way).

2.4 Pixel-level features

The following section describes the pixel-level features derived from multispectral and SAR images considered in this work (reported in Table 2.1) and their importance for monitoring crop growth. It was observed that choosing irrelevant features can lead to poor detection results, since unsupervised algorithms use all the features available for the analysis. For post-analysis and practical applications, it is also important to choose features whose interpretation is convenient in order to understand why an anomaly has been detected.

2.4.1 Multispectral vegetation indices

Many multispectral VIs have been introduced in the literature *e.g.*, (Bannari et al., 1995; Wu et al., 2008). A VI relates the acquired spectral information to the observed vegetation, and thus allows better quantitative and qualitative evaluations of the vegetation covers. The five multispectral VIs considered in this work are reported in Table 2.1 and described below. Note that raw S2 bands were also tested without any improvement in the detection precision and

a more difficult interpretation of the results when compared to VIs.

- **The NDVI** is a benchmark indicator for agronomic analyses and is mainly related to the plant vigor (Rouse et al., 1974; Bannari et al., 1995). Nowadays, the NDVI is still extensively used for vegetation monitoring (Klisch and Atzberger, 2016; Meroni et al., 2019; Garioud et al., 2021). This success is mainly due to the simplicity of NDVI computation, as well as its ability to efficiently capture information on vegetation phenology and health (Zeng et al., 2020).
- **The Normal Difference Water Index (NDWI)** actually refers to two different widely used indicators. The first version uses NIR and SWIR to monitor changes in the water content of leaves (Gao, 1996). The second version uses the green band and NIR to monitor changes related to content in water bodies (McFeeters, 1996). Both formulas are similar to NDVI with different bands involved. The SWIR version of NDWI seems to be more appropriate for crop analysis but the GREEN version of NDWI can also provide relevant information, *e.g.*, for flooded parcels.
- **The Modified Chlorophyll Absorption Ratio Index (MCARI)** was designed to extract information from the chlorophyll content in plants with a resistance to the variation of the Leaf Area Index (LAI). A variant called **MCARI/OSAVI** uses the Optimized Soil Adjusted Vegetation Index (OSAVI) to minimize the contribution of background reflectance (Daughtry et al., 2000; Wu et al., 2008).
- **The Green Red Vegetation Index (GRVI)** is similar to NDVI but uses the red and green bands. According to Motohka et al. (2010), GRVI "*can be a site-independent single threshold for detection of the early phase of leaf green-up and the middle phase of autumn coloring*" (referred to as senescence for crops).

2.4.2 SAR features

Many investigations have been performed to establish a relationship between SAR images and vegetation and have been reported in two recent reviews (McNairn and Shang, 2016; Liu et al., 2019). The backscattering coefficients (denoted as γ^0_{VH} and γ^0_{VV}) have been used intensively in the literature (Whelen and Siqueira, 2018; Khabbazan et al., 2019). The polarization ratio $\gamma^0_{\mathrm{VH}}/\gamma^0_{\mathrm{VV}}$, also used in various studies (Abdikan et al., 2016; Denize et al., 2018; Veloso et al., 2017; Vreugdenhil et al., 2018), was tested without showing any clear improvement. The same observation stands for the Radar Vegetation Index (RVI) (Kumar et al., 2013), which has been adapted to S1 with an alternative form $4\sigma^0_{VH}/(\sigma^0_{VH}+\sigma^0_{VV})$ (Nasirzadehdizaji et al., 2019). Thus, the results presented in this thesis have been obtained with the backscattering coefficients reported in Table 2.1.

Table 2.1: Pixel-level features computed from S2 and S1 images used in this work. 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.

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

2.5 Parcel-level features: Input data for the outlier detection algorithms

2.5.1 Extraction of parcel-level features with zonal statistics

The pixel-level features are averaged using spatial statistics referred to as "zonal statistics" in order to provide parcel-level features. Two zonal statistics are considered for the S2 VIs, namely the median and interquartile range (IQR). The median captures the mean behavior of a given parcel with more robustness than the classical mean as it is not affected by extreme values (Huber, 2011). It is used to detect anomalies affecting the entire agricultural parcel, such as anomalies in crop vigor. IQR is defined as the difference between the 75th and 25th percentiles. It contains information related to the heterogeneity of a given parcel while being robust to the presence of extreme values. These statistics were computed using the Python libraries SciPy version 1.4.1 (Virtanen et al., 2020) and rasterstats version 0.13.0[2]. Since cloud and shadow pixels were discarded, these statistics were computed from the remaining pixels after applying the cloud and shadow masks. Other zonal statistics were also tested, namely the skewness (which is related to the asymmetry of a distribution) and the kurtosis

[2] https://pythonhosted.org/rasterstats/, online accessed 8 December 2020

(which can be used to characterize the tail of a distribution) but led to a deterioration of the detection results (see results in Chapter 3). The set of SAR features reduces to the median of backscatter intensities, as IQR of S1 data is directly proportional to the median (more details regarding the choice of these statistics are provided in Section B.1.2).

2.5.1.1 Feature matrix

Each parcel is represented by a vector concatenating the zonal statistics computed for all pixel-level features at each date. The construction of the feature matrix, used as the input of the outlier detection algorithms, is illustrated in Table 2.2 when using the NDVI with 2 statistics. In the general case, the number of columns of this matrix is $N_{col} = N_{1,im} \times N_{1,f} \times N_{1,s} + N_{2,im} \times N_{2,f} \times N_{2,s}$, where $N_{1,im}$ 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 and similar definitions apply to $N_{2,im}$, $N_{2,f}$ and $N_{2,s}$ for S2 images. As each column corresponds to a unique combination statistics/feature/time, it is possible to compare each parcel columnwise. Finally, classical preprocessings such as the Principal Component Analysis (PCA) (Jolliffe, 1986) or the Multidimensional Scaling (MDS) (Borg and Groenen, 1997) were applied to this feature matrix without significant improvement regarding the outlier detection results. Thus, these preprocessing steps were ignored from our analysis.

Table 2.2: Simplified version of the feature matrix using NDVI only and two statistics (median/IQR) for n dates and M parcels. NDVI_{t_n} means NDVI computed for image $\#n$ and median_{P_M} means spatial median of the feature computed inside the parcel $\#M$

.

Parcel #	Feature 1	Feature 2	.	Feature L-1	Feature L
P_1	$\text{median}_{P_1}(\text{NDVI}_{t_0})$	$\text{IQR}_{P_1}(\text{NDVI}_{t_0})$	.	$\text{median}_{P_1}(\text{NDVI}_{t_n})$	$\text{IQR}_{P_1}(\text{NDVI}_{t_n})$
P_2	$\text{median}_{P_2}(\text{NDVI}_{t_0})$	$\text{IQR}_{P_2}(\text{NDVI}_{t_0})$	.	$\text{median}_{P_2}(\text{NDVI}_{t_n})$	$\text{IQR}_{P_2}(\text{NDVI}_{t_n})$
...	...	...	.	...	...
P_M	$\text{median}_{P_M}(\text{NDVI}_{t_0})$	$\text{IQR}_{P_M}(\text{NDVI}_{t_0})$	.	$\text{median}_{P_M}(\text{NDVI}_{t_n})$	$\text{IQR}_{P_M}(\text{NDVI}_{t_n})$

Chapter 3

Outlier detection at the parcel-level in wheat and rapeseed crops using multispectral and SAR time series

Part of this chapter has been adapted from the journal paper Mouret et al. (2021a).

Contents

3.1 Introduction

This Chapter presents two main contributions of this thesis. In a first step, a systematic description of the outlier parcels encountered throughout the study is conducted. In what follows, agricultural parcels are considered as abnormal (true positives) if they have an *agronomic* behavior significantly different from the majority of the other parcels. Errors in the parcel data (crop type reported or field boundaries) are also considered as true positives, since it is important to detect such problems in the database. On the other hand, noise or anomalies not relevant for crop monitoring (*e.g.*, undetected clouds) are considered as false positives since they are not relevant for the end user.

In a second step, outlier detection algorithms adapted to detect point anomalies are investigated and compared. We also provide complementary results obtained using the Isolation Forest (IF) algorithm to analyze the effect of changing various factors impacting the detection results. As mentioned in the introduction of this thesis, reducing the anomaly detection task to a point anomaly detection problem is a common approach since classical outlier detection algorithms can be used. Moreover, these approaches are compatible with the various constraints detailed in Chapter 1 (single growing season analysis, unlabeled dataset, etc). In that context, the feature matrix whose construction has been detailed in Chapter 2 is directly used as input of the outlier detection algorithm to find the most unusual samples.

Note that one potential limitation of point anomaly detection algorithms is that the temporal structure of the feature is not explicitly used (even if each column corresponds to a specific time instant). Methods using explicitly the time information of the data are investigated in Chapter 5.

3.2 Experiments conducted to evaluate the proposed method

In what follows, what is called "experiment" corresponds to an outlier detection conducted with a specific initial configuration (set of features, algorithm, outlier ratio, temporal interval) using one of the two datasets (wheat or rapeseed). Various experiments are conducted to evaluate the proposed approach: each time a new set of features or a new algorithm tuning was tested, the parcels declared as outliers were counterchecked by experts (if not previously detected), confirming the anomaly (true positive) or not (false positive), and determining the type of anomaly (see details later). This iterative procedure is illustrated in Figure 3.1.

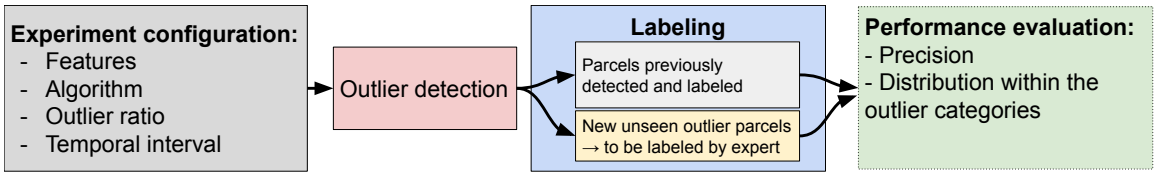


Figure 3.1: Diagram illustrating the idea behind the different experiments conducted.

For the rapeseed dataset, 252 initial configurations were tested to evaluate the factors that can influence the detection results. Most of these experiments were conducted on a complete growing season to evaluate the capacity of the proposed approach to detect anomalies occurring at different periods of the crop growth, and to determine whether differences between the detected parcels can be observed or not. Some other experiments were also made with a lower amount of data, in particular for a mid season analysis between October and February. Early detection can be of interest for warning purposes at the beginning of the growth cycle and gives more details on the effect of having only few images available for the analysis. The influence of the amount of parcels to be detected (called outlier ratio) is also tested to analyze the relevance of the outlier score given to each parcel.

For the wheat dataset, 25 experiments were made: the main idea was to determine whether our approach can be applied with minor modifications to other kinds of crops. The different experiments conducted during the study are reported in Table 3.1.

Table 3.1: Summary of the evaluated factors analyzed throughout the study. In parentheses, the number of different initial configurations tested (features, algorithm, time interval, outlier ratio).

Evaluated crop type	Time interval	Evaluated factors
Rapeseed (252)	Complete season (218)	Outlier detection algorithms Feature sets Outlier ratio Zonal statistics Missing S2 images Changes in parcel boundaries
	Mid season (34)	Feature sets, algorithms, outlier ratio
Wheat (25)	Complete season (20)	Feature sets, algorithm
	Mid season (5)	Feature sets, algorithm

3.3 Labeling and description of the outlier parcels

3.3.1 Outlier categories

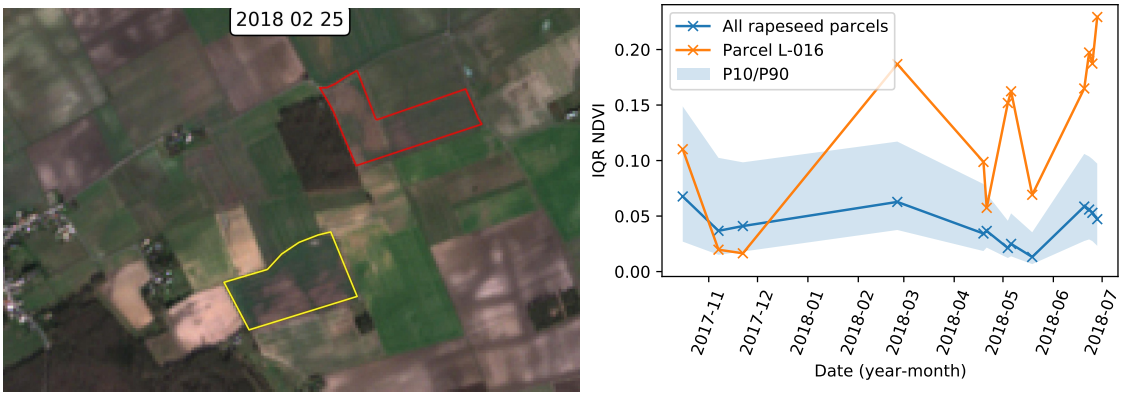
The outlier parcels were identified during multiple outlier detection analyses presented in Section 3.2. With the help of agronomic experts, the labeling of the detected parcels was conducted by visual-interpretation using all the available S1 and S2 images and by using all the time series of the different features/statistics to compare any analyzed parcel to the rest of the dataset. In order to compare the analyzed parcel to the rest of the data, the median, the 10th and the 90th percentiles of the whole dataset can be displayed (similar to a boxplot visualization). This representation allows the agronomic expert to easily know if the observed parcel has indicator values higher (or lower) than 90% of the data. Each detected parcel was then assigned to one of the outlier categories described in what follows.

The different anomalies analyzed throughout the study can be decomposed into 4 main categories: heterogeneity problems, growth anomalies, database errors and others. The category "others" corresponds to non-agronomic outliers that were considered not relevant for crop monitoring (referred to as false positives). A brief description of each category is proposed in Table 3.2 and more details and examples are provided below.

Table 3.2: Description of the different categories of anomalies detected during the labeling process. Subcategories were added to have a more precise description. For each category TP means true positive, considered relevant for crop monitoring, and FP means false positive, considered irrelevant for crop monitoring.

Category (TP/FP)	Subcategory	Description
Heterogeneity (TP)	Heterogeneity	Affects the parcel most of the season
	Heterogeneity (2 different parts)	The parcel is separated into two homogeneous different parts
	Heterogeneity after senescence	Occurs during senescence phase
	Early heterogeneity	Occurs during early growing season
Growth (TP)	Late growth	A late development is observed (non-vigorous crop)
	Vigorous crop	A vigorous development is observed
	Early flowers	Early flowering phase
	Early senescence	Early senescence phase
	Late senescence	Late senescence phase
Error in database (TP)	Wrong type	A wrong crop type is reported in the database
	Wrong shape	The parcel boundaries are not accurately reported
Others **(FP)**	Normal (checked)	The parcel was declared normal by the agronomic expert
	Too small	The parcel is too small, causing abnormal features
	SAR anomaly	Soil surface conditions causes abnormal SAR features
	Shadow perturbation (cloud or forest)	Shadows cause abnormality in the features.

- ***Heterogeneity*** corresponds to parcels presenting a clear heterogeneous development (i.e., spatially heterogeneous development). The most common cases of heterogeneity can be observed all along the growing season and are for instance related to soil heterogeneity, presence of weed or diseases. An example of a heterogeneous parcel is shown in Figure 3.2 (yellow boundaries). More transient cases of heterogeneity can affect the beginning (*early heterogeneity*) or the end of the growing season (*heterogeneity after senescence*) and can be for instance related to differences in soil characteristics or parcel exposure (e.g., Figure 3.4). *Heterogeneity (2 different parts)* parcels have two areas of the same crop separated by a clear frontier (e.g., strong difference in the phenological stages). An example of this type of anomaly is provided in Figure 3.3, where it can be observed that it is difficult to decide if wrong boundaries were provided, if two different varieties of rapeseed were sown or if soil differences led to heterogeneity.

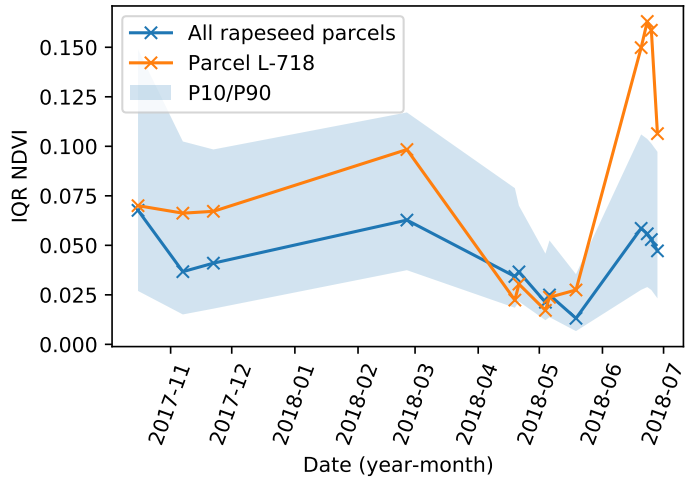


(a) (b)

Figure 3.2: Example of a heterogeneity affecting a parcel. (a): true color S2 image in February. (b): Interquartile range (IQR) of the parcel NDVI. 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 is the IQR NDVI for the analyzed parcel.

Figure 3.3: A rapeseed field (yellow boundaries) affected by a two-part heterogeneity. The left image was acquired in February 2018 and the right image in April 2018.

(a)



(b)

Figure 3.4: (a) A rapeseed field (yellow boundaries) affected by an heterogeneity after senescence. The left image was acquired in May and the right image in June. (b) Corresponding Interquartile Range (IQR) of the parcel NDVI time series. The blue line is the median value of the whole dataset. The blue area is filled between the 10th and 90th percentiles.

- ***Growth anomalies*** are related to an abnormal development of the crop. The two main categories of growth anomalies are parcels with a low vigor (*late growth*) or, on the contrary, with a high vigor (*vigorous crop*). Figure 3.5 illustrates how the different growth anomalies can affect the median NDVI of the parcels within a growing season. Figure 3.6 provides an example of growth anomaly where the S1 VH time series is affected by a late growth issue. Examples of vigorous parcels are provided in Figure 3.7. For wheat crops, we noticed that in the S2 image acquired in March 2017, a small amount of vigorous wheat parcels have the majority of their red pixels equal to zero, causing extreme values of the S2 features as illustrated in Figure 3.7. This issue was caused by the MAJA processing[1] and could be easily fixed using another processing chain like

[1] https://labo.obs-mip.fr/multitemp/using-ndvi-with-atmospherically-corrected-data/, online accessed 10 March 2020

Sen2core or with a threshold added to the red band. Since this phenomenon affected only a small amount of vigorous parcels, it did not change the quality of the detection results. As for heterogeneity, more transient growth anomalies, such as a delay in the flowering or senescence phase, can affect a parcel as illustrated in Figure 3.8.

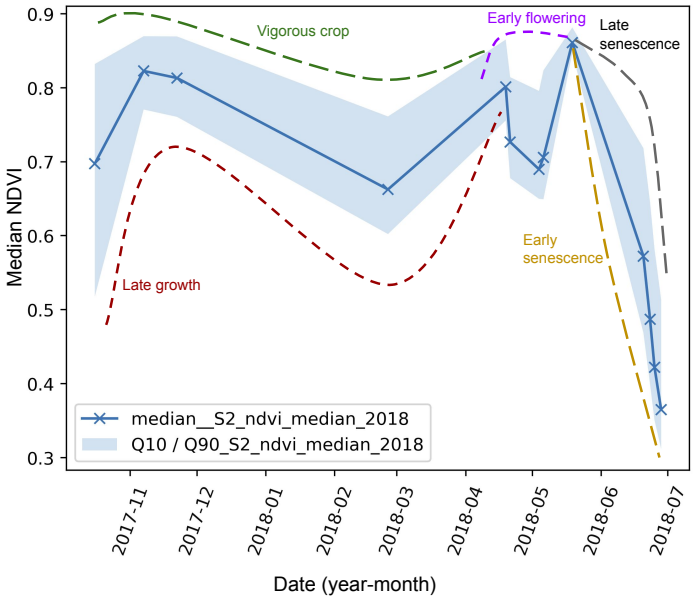
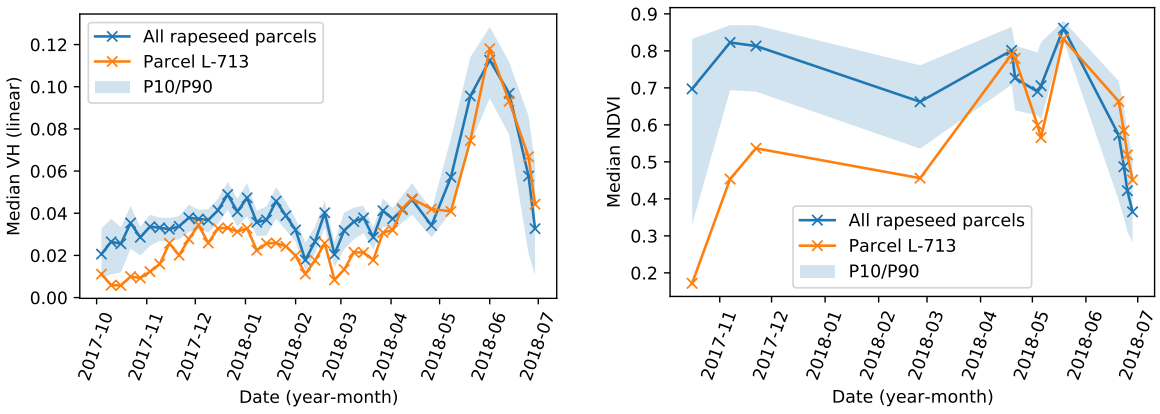
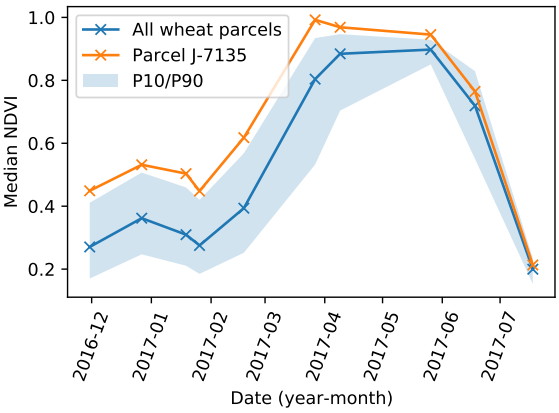


Figure 3.5: Illustration of the different growth anomalies that were detected and their potential influence on the median NDVI of the parcels (rapeseed crop). The blue line is the median value of the whole dataset. The blue area is filled between the 10th and 90th percentiles. Note that the labeling was conducted using all the S1 and S2 features (not only median NDVI).

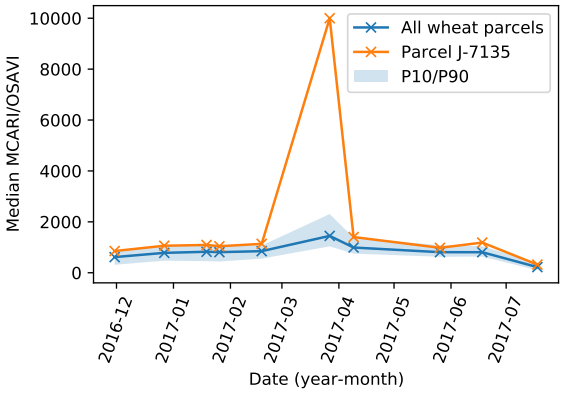


(a) (b)

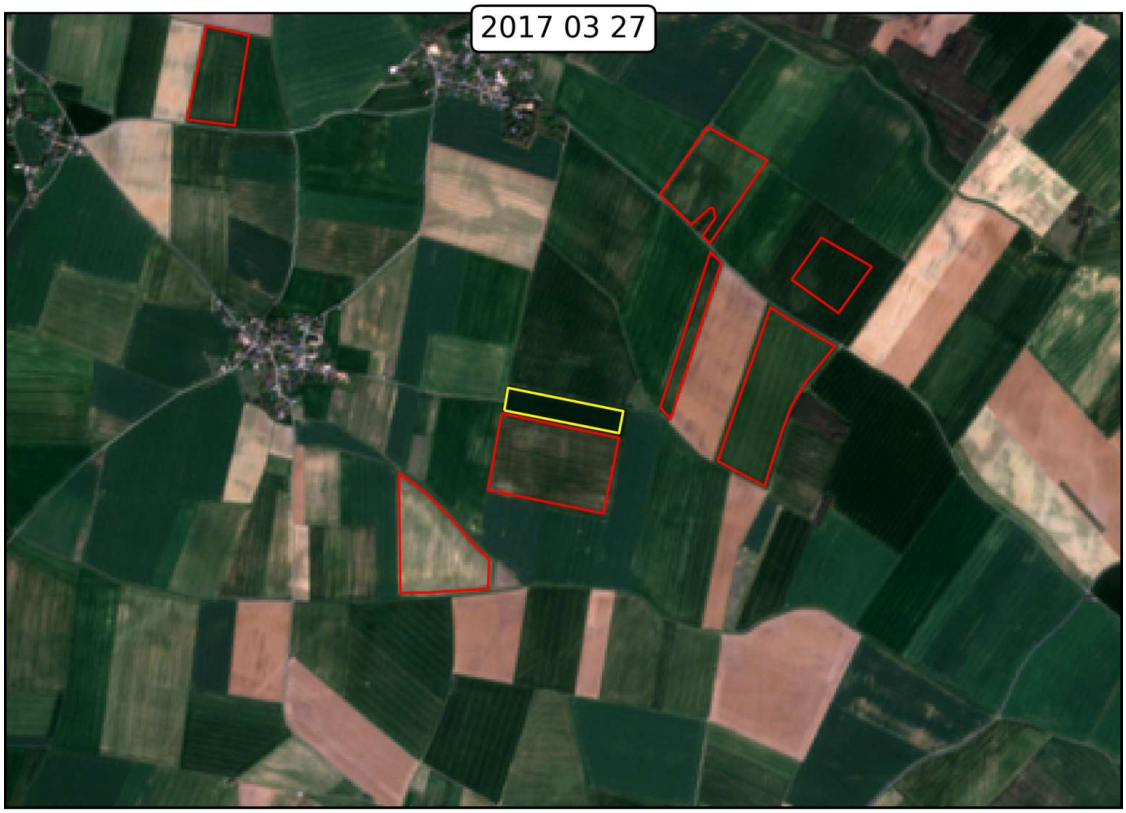
Figure 3.6: Example of time series subjected to late growth for a rapeseed parcel: (a) median VH and (b) median NDVI for a rapeseed parcel. The blue line is the median value of the whole dataset. The blue area is filled between the 10th and 90th percentiles. The orange line corresponds to a specific parcel subjected to late growth.



(a)

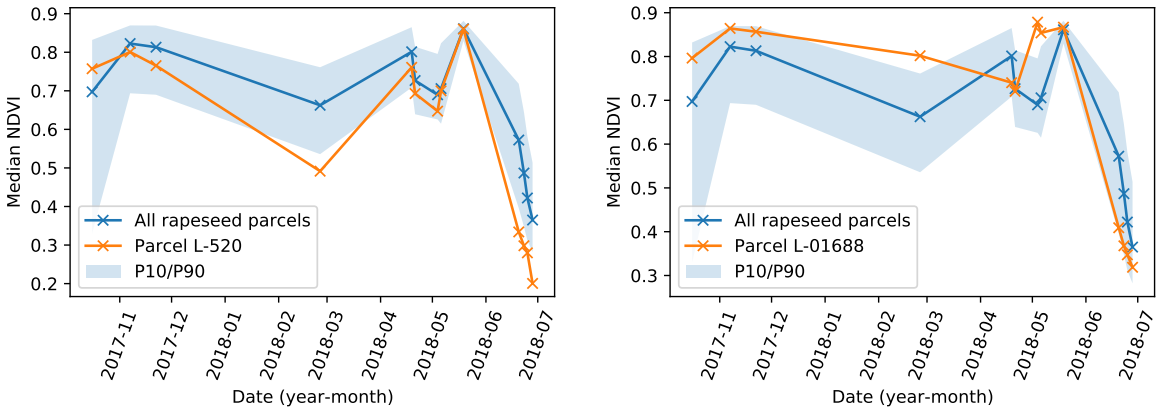


(b)



(c)

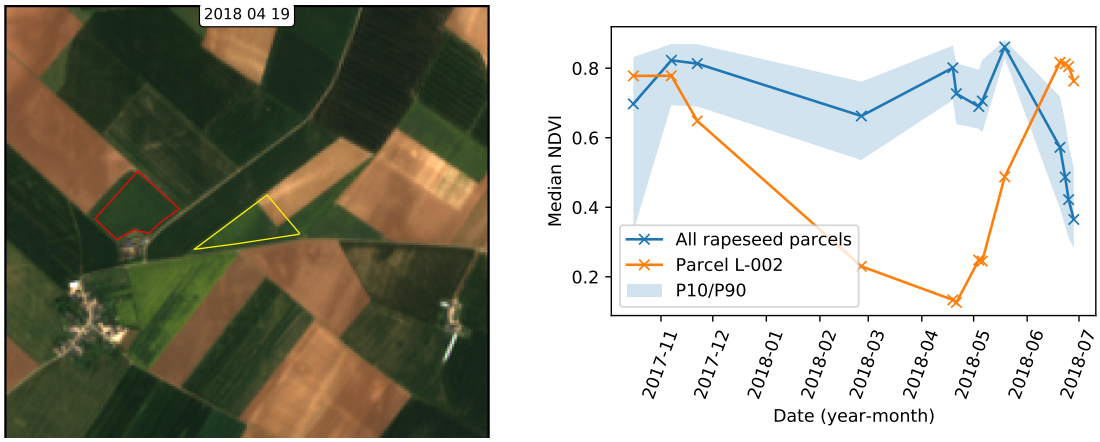
Figure 3.7: Example of time series subjected to a red channel problem in March 2017 for a wheat parcel: (a) median NDVI and (b) median MCARI/OSAVI (c) corresponding S2 image acquired in March (a parcel with a late growth can be observed at the bottom of the image (triangle with red boundaries)). The blue line is the median value of the whole dataset. The blue area is filled between the 10th and 90th percentiles. The orange line is a specific parcel analyzed.



(a) (b)

Figure 3.8: Time series of median NDVI for a rapeseed parcel presenting signs of (a) early senescence and (b) early flowering. 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 is a specific parcel analyzed.

- ***Database errors*** are considered as relevant anomalies to be detected. This type of error is a common problem in large databases and can be challenging and time consuming to be detected manually. Examples of "*wrong shape*" and "*wrong type*" reported in the database are provided in Figure 3.9. This category of anomalies presents in general a strong sign of abnormality.



(a) (b)

Figure 3.9: Two examples of error in the parcel contour database (a): an error in the parcel delineation is visible (true color S2 image). (b): median NDVI time series for a parcel having a wrong crop type declared.

- The ***"Normal (checked)"*** label was given to parcels that were labeled as normal after inspecting the features and images. In some cases, some few extreme values were observed explaining why the parcel was detected as abnormal by the outlier detection algorithms. In any case, all these parcels should have an outlier score (*i.e.*, the score given by an outlier detection algorithm) lower than the parcels affected by agronomic anomalies (*e.g.*, heterogeneity or growth anomaly).
- ***Other non-agronomic anomalies*** considered as false positives concern a small percentage of the analyzed parcels. Some very small sized parcels were still present in the dataset and are labeled as *"too small"* (it is sometimes difficult to clean efficiently too small parcels that are long and narrow). Analyzing this type of parcels is not possible due to the spatial resolution of Sentinel data. These parcels were kept in the database to illustrate problems that can occur in practical applications. *"Shadow"* is another kind of non-agronomic anomaly that can be caused by forests near the parcel (see Figure 3.10) or clouds that are not detected using the cloud mask.

Figure 3.10: Rapeseed parcels: the parcel with yellow boundaries is affected by shadow caused by the trees located next to the parcel. Also, at the bottom a too small parcel is visible.

- A subcategory of non-agronomic anomalies are ***"SAR anomalies"***. These anomalies correspond to parcels where SAR features have an abnormal time evolution in early growing season (*i.e.,* the SAR indicators are abnormal compared to the rest of the data), whereas multispectral images and their features were counterchecked as normal. It is a known issue in crop monitoring with SAR data that was studied in Wegmuller et al. (2006); Wegmüller et al. (2011); Marzahn et al. (2012), which is reported as a "Flashing field" phenomenon. These anomalies are considered as non-agronomic since SAR data are affected by other factors than the vegetation status such as soil moisture, soil structure, row orientation or soil roughness. This kind of anomalies was observed more frequently for wheat crops and in early growing season when there is a low vegetation cover. The "flashing field" terminology can easily be understood by looking at the example displayed in Figure 3.11.

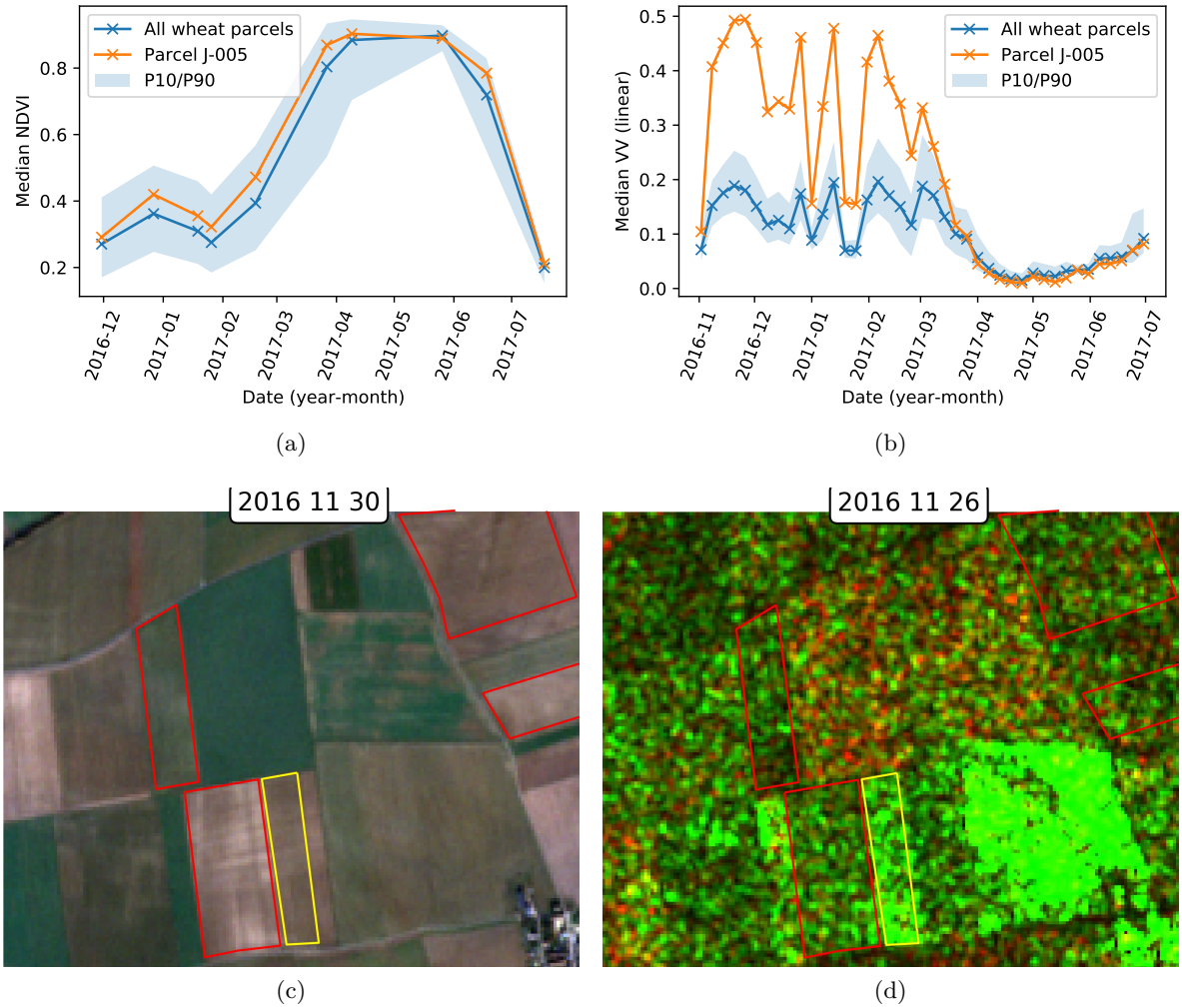


Figure 3.11: Time series of (a) median NDVI and (b) median VV polarization for a wheat parcel. The blue line is the median value of the whole dataset. The blue area is filled between the 10th and 90th percentiles. The orange line corresponds to a specific flashing-field parcel. Images acquired at the end of November: (c) true color S2 image and (d) S1 composite image (Green=VV, Red=VH).

3.3.1.1 Complementary information about SAR images and their anomalies

A strong correlation between SAR and plant vigor (late growth / vigorous crop) was observed in this study. Figure 3.12 illustrates the effect of late growth on S1 features. Figure 3.13 shows that SAR images are not always affected by heterogeneity within the parcel, as highlighted in the main document. Heterogeneity is detectable with SAR features when the crop structure is affected, which is understandable considering the nature of the sensor.

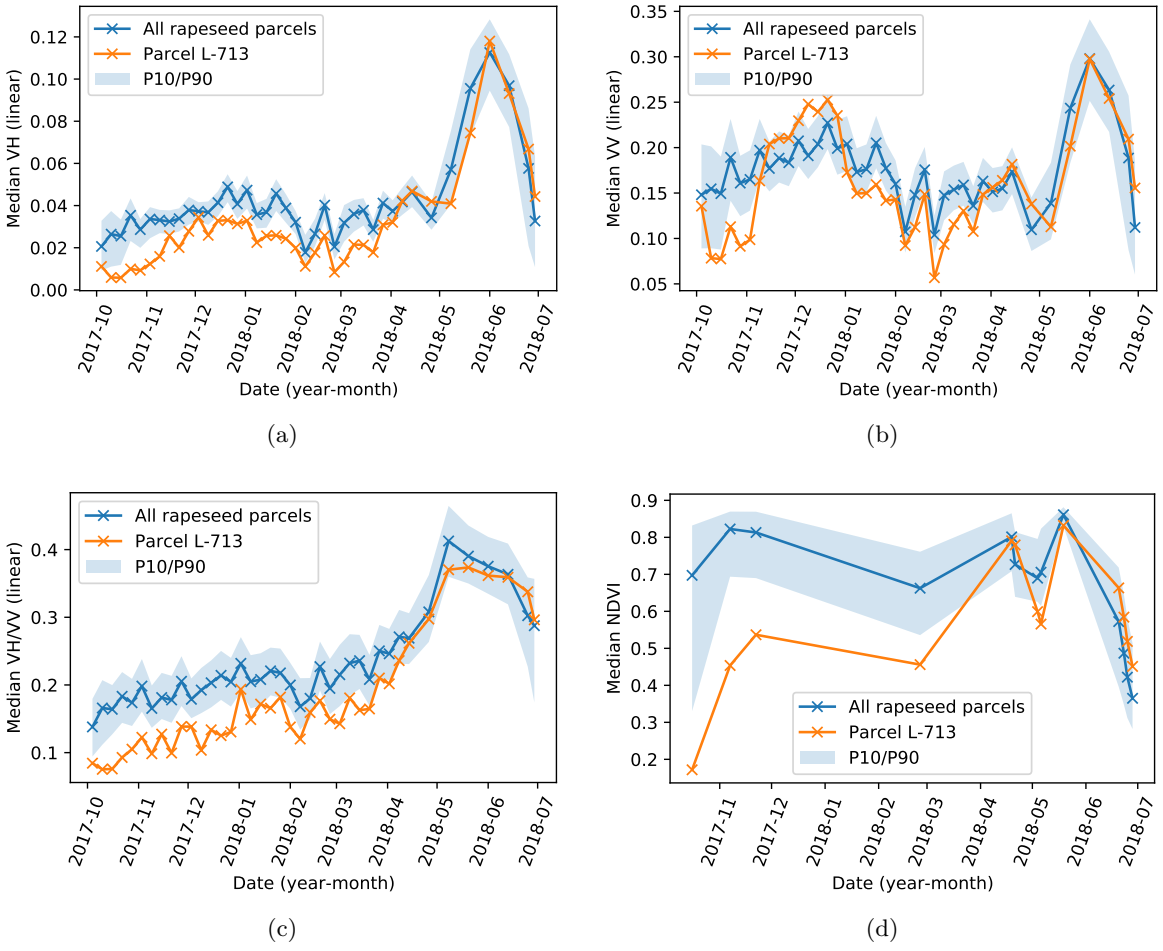


Figure 3.12: Time series of median SAR features (VV, VH, VH/VV) and median NDVI for a rapeseed parcel. The blue line is the median value of the whole dataset. The blue area is filled between the 10th and 90th percentiles. The orange line is a specific parcel analyzed.(a): median VH, (b): median VV, (c): median ratio VH/VV, (d): median NDVI

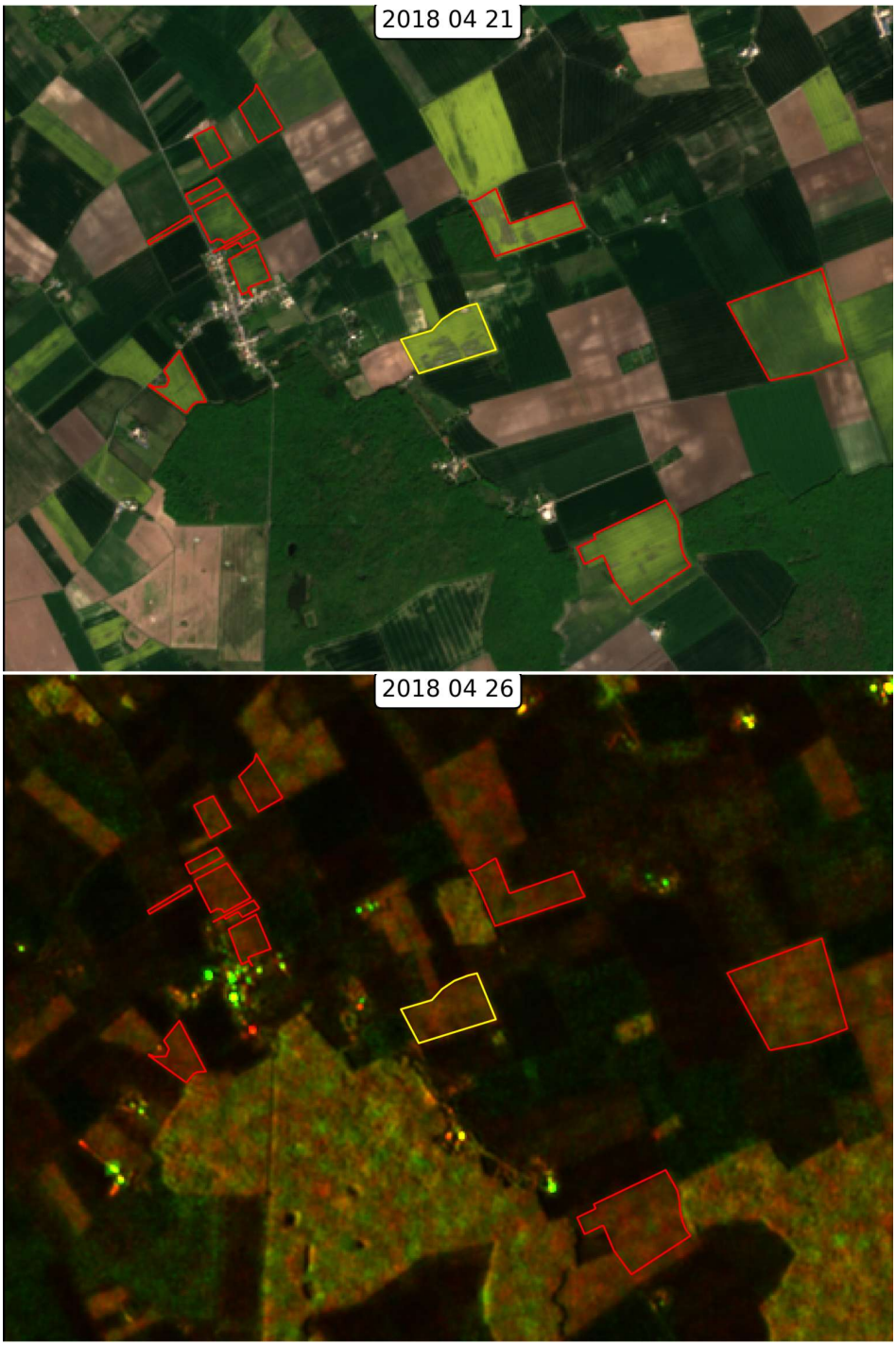


Figure 3.13: Example of a parcel of rapeseed crop (yellow boundaries) where heterogeneity occurs almost during the complete season. Some other parcels show some signs of heterogeneity too. Top: true color S2 image acquired in May, bottom: composite SAR image (green channel is VV polarization and red channel is VH polarization) acquired in May with multi-temporal speckle filtering.

3.3.2 Distribution of the labeled parcels in the two datasets

Figure 3.14 summarizes the distribution of the anomaly categories for both wheat and rapeseed crops. Approximately 55% of the rapeseed dataset was checked by the agronomic experts, ensuring that the outlier parcels analyzed in the study are representative. Similarly, 30% of the most abnormal wheat parcels were checked to validate the relevance of our method when applied to another crop type. Figure 3.14 shows that heterogeneity and growth problems are the most detected anomalies for both types of crops.

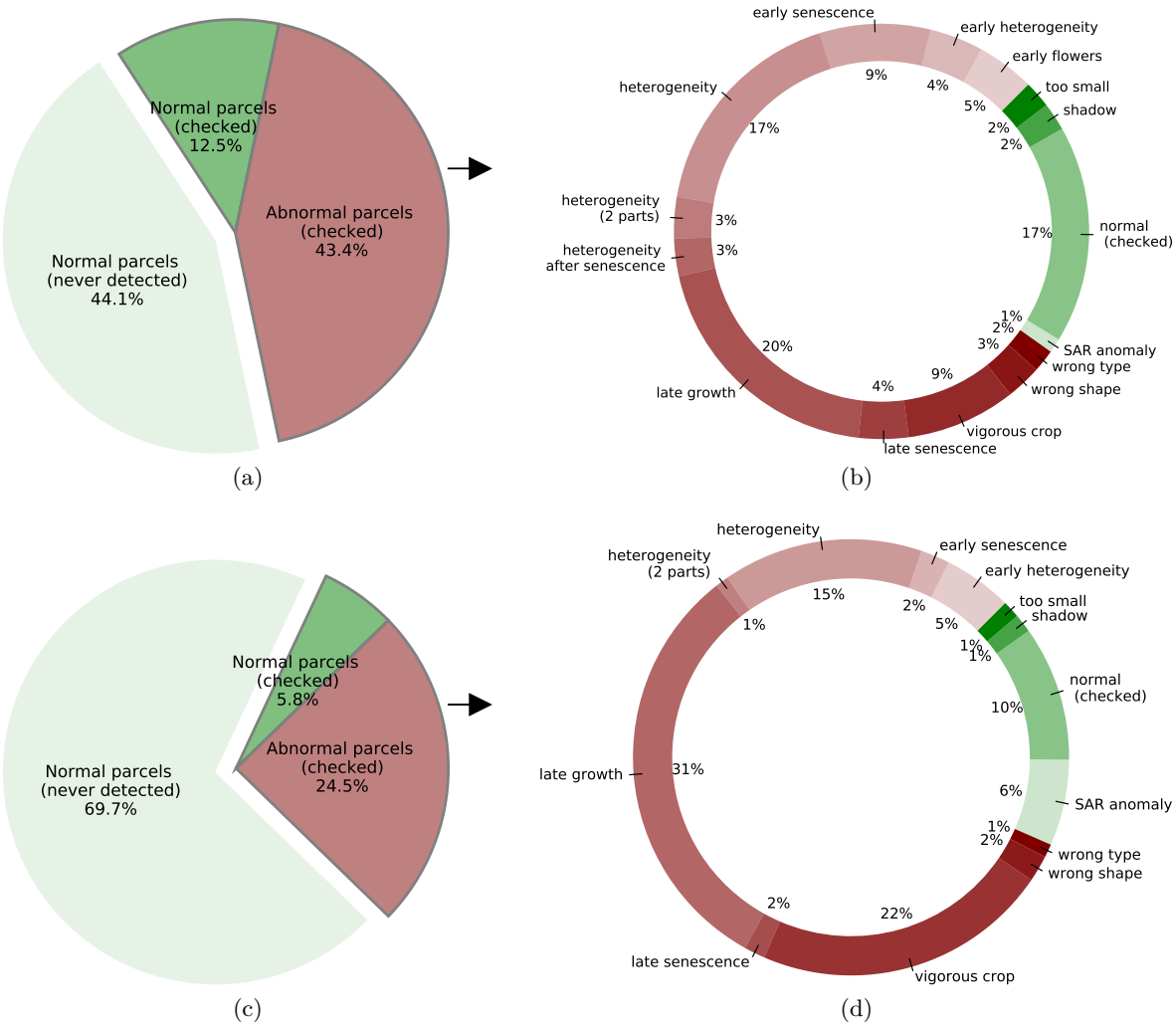


Figure 3.14: Distribution and description of the labels of the parcels for (a, b) rapeseed crops and (c,d) wheat crops. Red categories correspond to abnormal parcels that have been labeled and categorized by experts. Green categories correspond to normal parcels and are divided in 2 main groups in (a) and (c): 1) the normal parcels never detected during the conducted experiments that have not been checked by experts and 2) the normal parcels that were detected during the experiments and have been declared normal by experts.

3.4 Performance evaluation for quantifying the quality of the detection results

The precision is used to evaluate the quality of a detection and is defined as

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.1}$$

where TP and FP are the numbers of true positives and false positives, respectively. The precision expresses the percentage of detected parcels that are true positives (here, agronomic anomalies checked by the experts). Plotting precision vs. outlier ratio curves is a good way to compare various detection results: for a given outlier ratio, a good algorithm or feature choice has generally detection results with a higher precision. Note that these curves also provide information regarding the false negatives samples since for a given outlier ratio, a higher precision means less false negatives. These curves are similar to the Receiver Operating Characteristics (ROC) or the precision vs. recall curves but with the advantage of being more adapted to outlier detection (Saito and Rehmsmeier, 2015). Indeed, the outlier ratio can be adjusted without ground-truth, by selecting the parcels with the highest outlier scores. Moreover, when analyzing these curves one can focus on realistic values of the outlier ratio (*e.g.*, precision obtained when detecting more than 50% of the data instances seems not adapted to our problem). More details on these evaluation curves are provided in Appendix A. The area under the precision vs. outlier ratio curve (AUC) can be used to provide a quantitative measure of detection performance summarizing the information contained in the whole curve. In the analysis, we computed the AUC for outlier ratios in the range [0, 0.5]. The AUC was then divided by 0.5 to normalize the obtained value: the resulting score can be seen as the average precision for outlier ratios in the range [0, 0.5].

This representation does not give information regarding the distribution of the different detected categories since two algorithms can have the same precision without detecting the same parcels (*e.g.*, one algorithm can detect more heterogeneous parcels whereas another one detects more late growth anomalies). Using the distribution of the different types of anomalies detected for a given outlier ratio is a complementary way to address this limitation. Note that to highlight the distribution of the false negative instances, it is possible to display the number of detected parcels in each category divided by the total number of parcels in each category.

3.5 Comparing unsupervised outlier detection techniques for crop monitoring at the parcel-level

3.5.1 Outlier detection algorithms

This section provides a general reminder about the four benchmark algorithms tested in this work. Implementation details and hyperparameter tuning related to our specific use case are provided in Section 3.5.2.1.

3.5.1.1 Local Outlier Probabilities

The LoOP algorithm (Kriegel et al., 2009) is based on the nearest neighbors of the observed samples. It is a probabilistic extension of the local outlier factor (LOF) algorithm (Breunig et al., 2000). The main idea behind LoOP is that normal data instances occur in dense neighborhoods and that anomalies occur far from their closest neighbors (Chandola et al., 2009). The LoOP algorithm is briefly detailed in what follows. Let knn($\boldsymbol{P}$) be the set of k nearest neighbors of a sample $\boldsymbol{P}$. Let $d(\boldsymbol{P}, \boldsymbol{O})$ be the distance between the object $\boldsymbol{P}$ and $\boldsymbol{O}$ (*e.g.*, the Euclidean distance is generally used). The LoOP algorithm first introduces a (local) *probabilistic distance* denoted as *pdist*, which aims to be less sensitive to the choice of k:

$$\text{pdist}(\lambda, \boldsymbol{P}, \text{knn}(\boldsymbol{P})) = \lambda \sqrt{\frac{\sum_{\boldsymbol{O} \in knn(\boldsymbol{P})} d(\boldsymbol{P}, \boldsymbol{O})^2}{|\text{knn}(\boldsymbol{P})|}} = \lambda \sqrt{\mathbf{E}_{\boldsymbol{O} \in \text{knn}(\boldsymbol{P})} \left[d(\boldsymbol{P}, \boldsymbol{O})^2 \right]} \tag{3.2}$$

where the expected value of a random variable $\boldsymbol{X}$ is denoted $\mathbf{E}[X]$ and λ is known as the "extent" parameter (Constantinou, 2018). This parameter defines the statistical notion of an outlier as an object deviating more than a given λ time the standard deviation from the mean. For example, λ=2 implies outliers deviating more than 2 standard deviations and correspond to 95% in the empirical three sigma rule. Note that *pdist* can be seen as an estimation of the density around $\boldsymbol{P}$.

In a second step, Probabilistic Outlier Factor (PLOF) and its standard deviation (nPLOF) are computed:

$$\text{PLOF}_{k,\lambda}(\boldsymbol{P}) = \frac{\text{pdist}(\lambda, \boldsymbol{P}, \text{knn}(\boldsymbol{P}))}{\frac{\sum_{\boldsymbol{O} \in knn(\boldsymbol{P})} \text{pdist}(\lambda, \boldsymbol{P}, \boldsymbol{O})}{|\text{knn}(\boldsymbol{P})|}} - 1 = \frac{\text{pdist}(\lambda, \boldsymbol{P}, \text{knn}(\boldsymbol{P}))}{\mathbf{E}_{\boldsymbol{O} \in \text{knn}(\boldsymbol{P})} \left[\text{pdist}(\lambda, \boldsymbol{P}, \boldsymbol{O}) \right]} - 1 \tag{3.3}$$

$$\text{nPLOF} = \lambda \sqrt{\mathbf{E}\left[\text{PLOF}^2 \right]} \tag{3.4}$$

PLOF is similar to the LOF score (Breunig et al., 2000) and nPLOF can be seen as the standard deviation of PLOF values, assuming mean(PLOF)=0. The LoOP score is finally

computed using the Gaussian Error Function (noted "erf"):

$$\mathrm{LoOP}_{k,\lambda}(\boldsymbol{P}) = \max\left(0, \mathrm{erf}\left(\frac{\mathrm{PLOF}_{k,\lambda}(\boldsymbol{P})}{\mathrm{nPLOF}\sqrt{2}}\right)\right) \tag{3.5}$$

This final normalization step allows LoOP to provide a score in the range [0,1], which is consistent in every region of the dataset. An example of the LoOP scores obtained in a 2 dimensional dataset is provided in Figure 3.15 extracted from Kriegel et al. (2009).

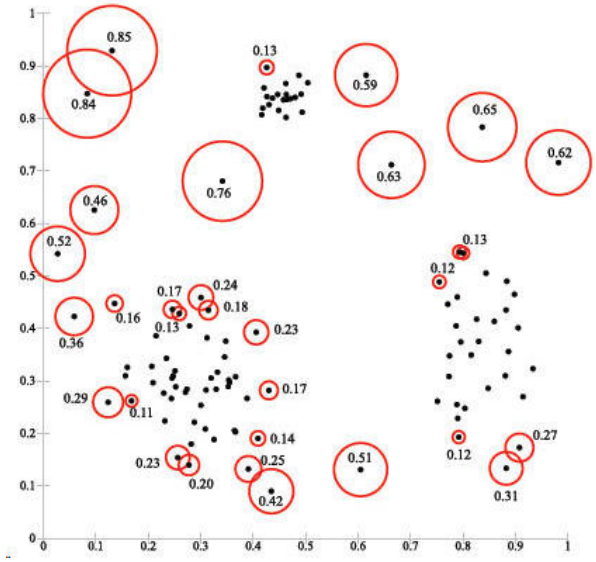


Figure 3.15: LoOP values on 2D synthetic data, with $k = 20$ and $\lambda = 3$ (Kriegel et al., 2009).

The Python library PyNomaly (version 0.3.3) was used for the implementation of the LoOP algorithm (Constantinou, 2018). Two hyperparameters have to be fixed: k, the number of nearest neighbors and the extent parameter λ.

3.5.1.2 Autoencoders

AE have been considered intensively for feature learning and dimensionality reduction (Kramer, 1991) and have been popularized thanks to the advent of deep learning. Similarly to other dimensionality reduction techniques such as PCA, AE can be used for outlier detection: the idea is that outliers tend to have a larger reconstruction error compared to nominal vectors (Aggarwal, 2017). AEs are able to learn a non-linear representation of the data for classification or outlier detection. However, they tend to be subject to overfitting and convergence issues. A classical autoencoder structure is depicted in Figure 3.16.

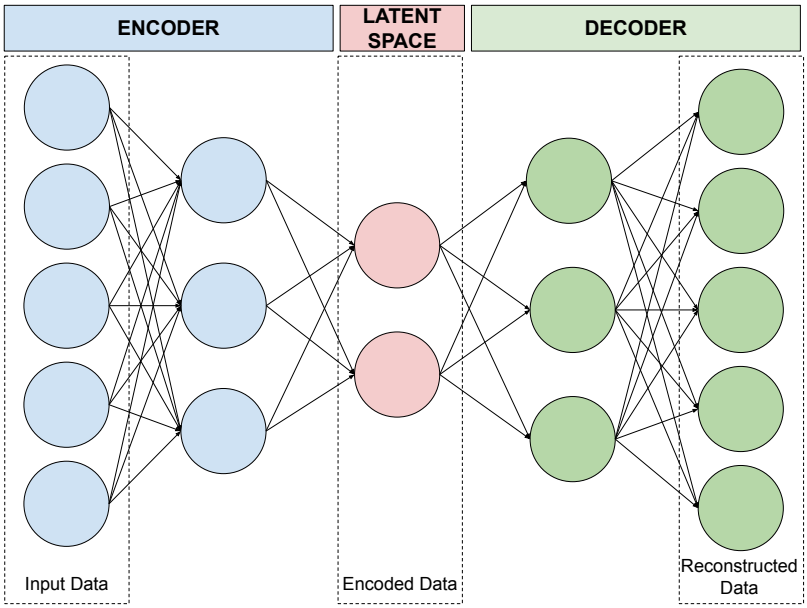


Figure 3.16: Schematic picture of an autoencoder architecture.

The implementation of the AE algorithm was made using the Python library Keras[2] (version 2.3.0). AEs need a large amount of parameter tunings to work efficiently, *e.g.*, the user needs to set the number of hidden layers, the activation functions, the regularization parameters, the loss function, the number of epochs for training, the batch size, etc.

3.5.1.3 Isolation Forest

The IF algorithm (Liu et al., 2012) aims at detecting anomalies without using any distance or density measure by assuming that outliers can be isolated more easily than other instances. Using binary isolation trees to separate instances, outliers are more likely to be isolated at the root of the trees whereas inliers tend to be isolated at deeper parts of the trees as illustrated in Figure 3.17. The IF algorithm constructs multiple random isolation trees defining a so-called forest of iTrees. The construction of an iTree is random: at each node, a random feature is chosen with a random split value. When using random splits with random features, outliers are more likely to be isolated first. The number of splitting required to isolate an instance is called the path length. The anomaly score of a given instance can be defined from the averaged path length in the forest. Outliers tend to have a short average path length whereas inliers are isolated with a large number of splits. The IF algorithm is known to be very fast compared to other algorithms (especially for large datasets), since for instance comparing a number to a threshold does not need to calculate complicated test statistics. Using a large number of iTrees generally improves convergence and makes the algorithm less sensitive to the random nature of the trees.

[2] `https://keras.io/`, online accessed 04 February 2021

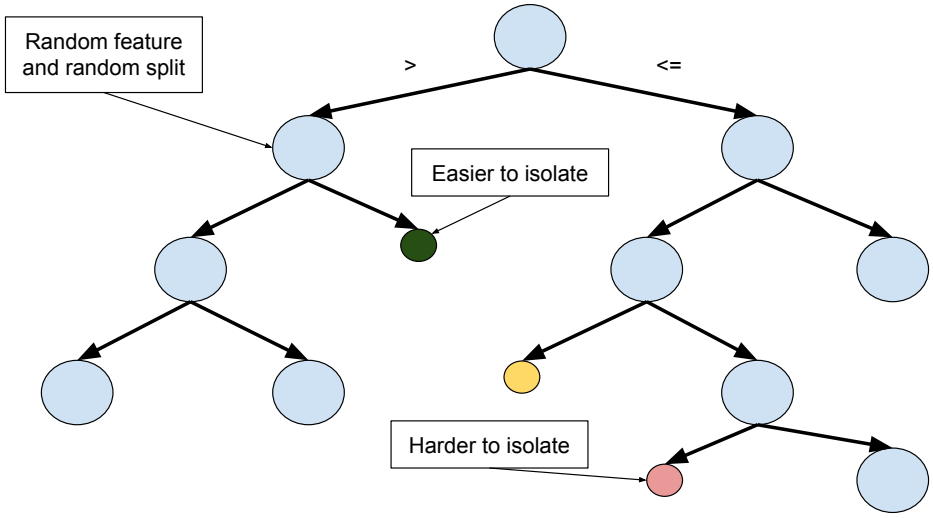


Figure 3.17: Isolation tree: outliers tend to be isolated much faster than inliers.

The Python library Scikit-learn (version 0.23.0) was used for the implementation of IF (Pedregosa et al., 2011). Hyperparameters that have to be tuned are the number of iTrees n_{trees} and the size of the data subsampling $n_{samples}$ used to construct the iTrees.

3.5.1.4 One-Class Support Vector Machine

OC-SVM (Schölkopf et al., 1999) is a model-based technique assuming that normal instances of the training set are part of the same class delimited by a separating boundary (Chandola et al., 2009). The instances that are not inside this boundary are then considered as anomalies. OC-SVM, as defined in Schölkopf et al. (1999), determines the maximal margin hyperplane between the data points and the origin. For a set of instances $\boldsymbol{x}_i \in \boldsymbol{X}$, with a separating hyperplane defined by $w^T\boldsymbol{x} + b = 0$, the OC-SVM algorithm solves the following problem:

$$\min_{w,\xi,\rho} \frac{1}{2}||w||^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho, \tag{3.6}$$

subject to

$$w^T\phi(\boldsymbol{x}_i) \geq \rho - \xi_i \quad \text{and} \quad \xi_i \geq 0, \forall i = 1, ..., n. \tag{3.7}$$

The hyperparameter ν is an upper bound for the fraction of training samples located outside the frontier that has to be fixed by the user (it is the amount of outliers to be detected even if this number is not guaranteed). The variables ξ_i are slack variables, which allow the classifier to create a soft margin in order to avoid overfitting. Finally, using a kernel associated with the non-linearity ϕ transforms the OC-SVM linear model into a non-linear model. The radial basis function (RBF) kernel K between two vectors $\boldsymbol{x}$ and $\boldsymbol{x}'$ is defined as follows

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\right). \tag{3.8}$$

The output of OC-SVM is deterministic for a given training set since it is the solution of a convex optimization problem. Moreover, the OC-SVM algorithm has a particularity regarding the anomaly score given to each instance. The three other algorithms tested here provide an anomaly score to each instance independently from the amount of outliers to be detected, called the outlier ratio. Conversely, in order to construct the OC-SVM boundary, the outlier ratio ν has to be fixed by the user. If the percentage of anomalies to be detected is changing with time, the separating frontier needs to be updated accordingly. The choice of ν can significantly impact the behavior of the classifier (Schölkopf et al., 1999).

The Python library Scikit-learn (version 0.23.0) was used for the implementation of OC-SVM (Pedregosa et al., 2011). An RBF kernel was investigated, motivated by its effectiveness that has been observed in many applications (Schölkopf et al., 2004). The RBF kernel has a single hyperparameter σ referred to as kernel bandwidth, which has to be adjusted for each dataset.

3.5.2 Comparison results on rapeseed and wheat crops

This section analyzes the performance of each algorithm applied to crop monitoring. Obviously, the outlier detection algorithm should be able to detect a majority of relevant anomalies. For practical reasons the algorithm with the simplest and most robust hyperparameter tuning should be preferred. Finally, the algorithm should also provide results that are stable for a given configuration and robust to changes in the feature set and crop type.

3.5.2.1 Hyperparameter tuning

Hyperparameter tuning is an important step in the design of outlier detection algorithms. As explained in (Aggarwal, 2017, Section 13.10.1), having an outlier detection algorithm whose results are highly dependent on the choice of its parameters can lead to poor results when applied to a broad range of real-world datasets. This section provides details about our hyperparameter tuning and their influence on the different results. This tuning was conducted using the rapeseed dataset, and tested without any change using the wheat crop to analyze the robustness to crop changes. All the hyperparameters used in the study are reported in Table 3.3.

For the LoOP algorithm, the number of nearest neighbors k was fixed by grid search leading to $k = 701$. This value provided detection results of higher precision compared to the other tested values (small changes in the value of k do not significantly affect the results). It was found that choosing a too small number of neighbors (*e.g.*, choosing an odd-valued integer close to the square root of the number of observations as proposed in Constantinou (2018) leads to detect too subtle anomalies that are not related to agronomic issues. Intuitively, choosing a relatively high number for k means that anomalies are defined with respect to the majority of the data, which seems coherent when looking at the behavior of the abnormal parcels. Nevertheless, this sensitivity to the number of neighbors, discussed for instance in

Aggarwal (2017, Section 13.10.1) can be problematic when changing the dataset. The extent parameter of LoOP was fixed to $\lambda = 2$ as recommended in Constantinou (2018). The value of λ did not have a significant influence on the detection results for the rapeseed and wheat crops.

For the OC-SVM algorithm, an efficient heuristic ((Jaakkola et al., 1999), (Aggarwal, 2017, p.93)) consists of estimating the parameter σ as the median of the pairwise Euclidean distances between vectors from the learning set $\boldsymbol{X}$, denoted as median(dist($\boldsymbol{X}$)). This estimator of σ provided good results without a need for a manual tuning for each new dataset.

The parameters of the AE were tuned by grid search. We considered a classical structure similar to the one proposed in the Python library for outlier detection PyOD (Zhao et al., 2019): 4 hidden layers with 64, 32, 32 and 64 neurons. A Relu activation function was used for all layers except for the output using a sigmoid function. Layer weights were regularized using an ℓ_2 penalty with a regularization value (referred to as "kernel regularizer" in Keras) set to 10^{-4}. This specific regularization significantly improved the detection results, contrary to changes in the network structure (e.g., number of neurons). In particular, a small regularization induces some overfitting, making the separation between inliers and outliers difficult since the reconstruction error is close to zero for every sample. On the other hand, when the regularization is too strong, the AE tends to reconstruct all the time series by a simple linear regression, which is clearly not satisfactory. Another important hyperparameter is the number of epochs, which has to be fixed to avoid underfitting or overfitting. Since the rapeseed and wheat datasets are relatively small, 10 epochs were sufficient to obtain good detection results with a batch size of 128 samples. Considering the large number of parameters to be set and their influence on the behavior of the algorithm, it is clear that AE is the most challenging algorithm to use in practice when compared to the other tested algorithms. For an unsupervised task, this can be problematic, as explained in the introduction of this Section.

The IF algorithm was used with a number of iTrees equal to $n_{\text{trees}} = 1000$ and a subsampling fixed to $\text{n}_{\text{samples}} = 256$ as in the original paper (Liu et al., 2012). Changing these two parameters did not have a significant effect on the results, which is an advantage compared to the other algorithms. This robustness of the IF algorithm with respect to hyperparameter tuning is coherent with the observations made in (Aggarwal, 2017, Section 13.10.1).

Because they are sensitive to scaling, the OC-SVM, LoOP and AE algorithms also require a normalization step in order to have input features in the interval $[0, 1]$, while this step is not mandatory when using the IF algorithm.

Table 3.3: Hyperparameters used in the different algorithms

Algorithm	Hyperparameter	Value
IF	n_{trees}	1000
	$n_{samples}$	256
LoOP	k	701
	λ	2
AE	hidden neurons	64, 32, 32, 64
	epoch	10
	output regularization	10^{-4}
OC-SVM	σ	median(dist($\mathcal{X}$))

3.5.2.2 Performance evaluation

In practical applications, the percentage of parcels to be detected or analyzed can depend on the user needs. Thus, it is important to evaluate the performance of the algorithms for different outlier ratios, as explained in Section 3.4. This outlier ratio corresponds to parameter ν in OC-SVM, to the $(1-\nu)\%$ highest anomaly scores in LoOP, to the $\nu\%$ highest reconstruction errors in the AE algorithm, or to the $\nu\%$ highest average path length for IF. Recall that precision versus outlier ratio curves summarize the detection performance for all the outlier ratios. They provide similar information as the receiver operating characteristics (ROC) curves that are classically used in detection theory, with the advantage of being more adapted to outlier detection since the classes are unbalanced (Saito and Rehmsmeier, 2015).

Precision versus outlier ratio curves (averaged using 100 Monte Carlo runs) obtained for the rapeseed dataset are displayed in Figure 3.18 when using the features detailed in Chapter 2. All the tested algorithms reach similar precision for a given outlier ratio, showing that the multiple methods for detecting anomalies provide consistent results. However, the IF algorithm provides the best overall performance with an area under the curve AUC = 0.885. The better performance of IF can be particularly observed for high outlier ratios (higher than 0.3), allowing more subtle anomalies to be detected.

LoOP and OC-SVM algorithms provide a unique solution for a given dataset, contrary to AE and IF whose results vary from one run to another. In order to evaluate the variability of the results for these two algorithms, 100 Monte Carlo simulations have been performed. The distributions of the resulting AUC values obtained for the precision vs. outlier ratio curves are displayed in Figure 3.19. The AE algorithm provides less stable results when compared to IF, with a minimum AUC close to 0.83 (minimum AUC is 0.88 for IF). Based on the analysis, the use of the IF for outlier detection in agricultural parcels is recommended.

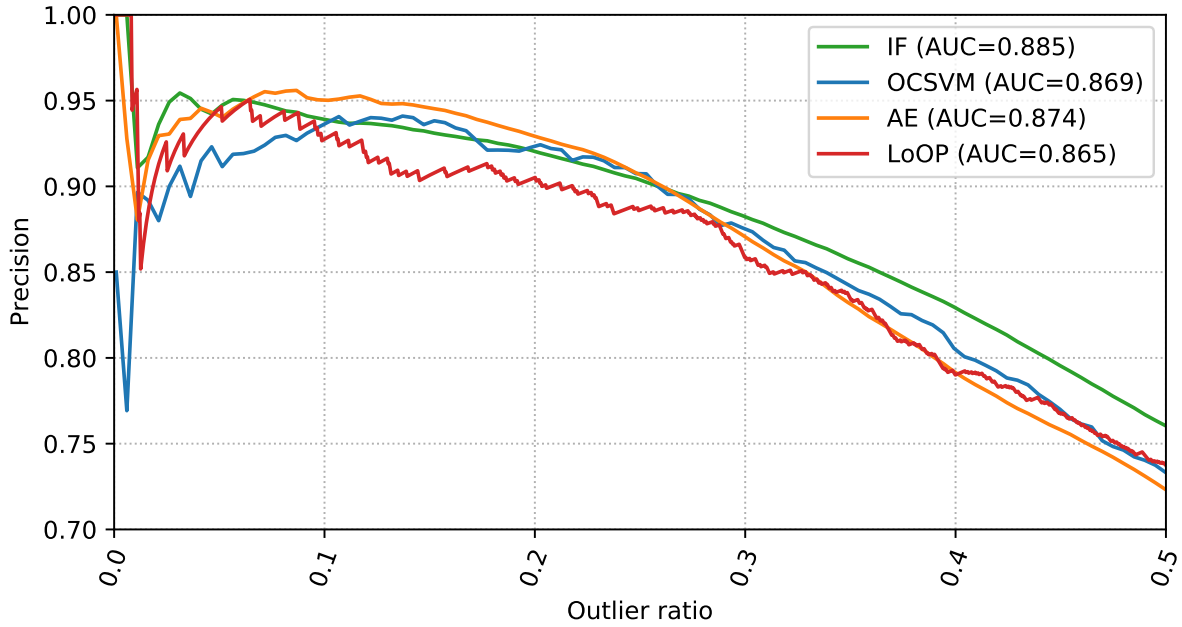


Figure 3.18: Precision vs. outlier ratio curves for the rapeseed dataset (averaged using 100 Monte Carlo runs). AUC means area under the curve computed for outlier ratios in the range $[0, 0.5]$ (i.e., the average precision in that range).

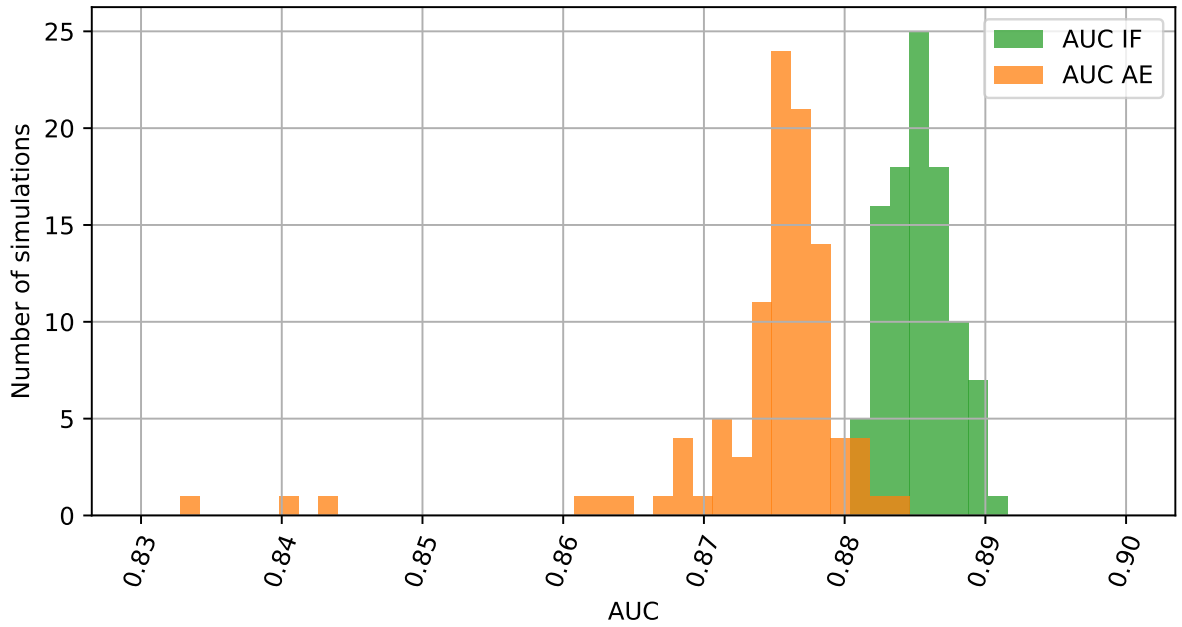


Figure 3.19: Distribution of the area under the precision vs. outlier ratio curves obtained on the rapeseed dataset (100 iterations).

3.5.2.3 Sensitivity to a crop change

This section analyzes the robustness of the algorithm with respect to other types of crops using the wheat dataset. The different algorithms were run with the hyperparameters determined for the rapeseed dataset. The precision of the results obtained for an outlier ratio $\nu = 0.10$ (which is a realistic choice from an operational point of view) is reported in Table 3.4. OCSVM and LoOP seem to be more affected by a crop change, when compared to AE and IF. For OCSVM and LoOP, a significant improvement was observed after a good hyperparameter tuning, confirming instability with respect to the choice of the parameters (precision after tuning is 92.80 for LoOP and 91.3 for OCSVM).

Table 3.4: Precision of the results with an outlier ratio fixed to 10%

Crop type	IF	AE	LoOP	OCSVM
Rapeseed	94.1	95.0	92.7	93.6
Wheat	95.5	95.8	86.0	89.1

3.5.3 Conclusions regarding the outlier algorithm to be used

Overall, the IF algorithm provided the best performance: a precision of 95% is reached for both rapeseed and wheat crops for an outlier ratio of 10%. A similar precision is obtained using autoencoders. However this technique is more difficult to implement due to its large number of hyperparameters. Moreover, it provides more instability when considering different initializations. The one-class SVM and local outlier probabilities algorithms provided similar results for the rapeseed crops, but did not scale well to a change in the dataset. This lack of robustness can be a problem in practical applications. As a consequence, the IF algorithm seems to be well adapted to crop monitoring at the parcel-level.

3.6 Detailed results on rapeseed and wheat crops using the Isolation Forest algorithm

The different feature combinations tested in this section are identified in the figures using abbreviations that are defined in Table 3.5.

Table 3.5: Abbreviations used with their corresponding sets of features used for outlier detection. Each abbreviation can be read as follows: "sensor: pixel-level feature (parcel-level statistics)".

Abbreviated name	Features used
S1: VV, VH (median)	Median of S1 features listed in Chapter 2
S2: all (median / IQR)	Median and IQR of all S2 features listed in Chapter 2
S2: all (median / IQR), S1: VV, VH (median)	Median and IQR of all the S2 features and median of the 2 S1 features VV and VH.

3.6.1 Anomaly detection results for rapeseed crops

The results presented in this section were conducted by analyzing the complete rapeseed dataset with the IF algorithm. First, the outlier detection is conducted using S1 features only, since SAR data are available permanently through all the crop cycle, which is important for crop monitoring applications. Then, the effect of using S2 features only is investigated. Finally, S1 and S2 features are used jointly to study the effect of combining the contribution of both sensors.

3.6.1.1 Outlier detection with S1 features

The strength of S1 data for crop anomaly detection is confirmed when analyzing Figure 3.20 (black curve): the precision is equal to 92.3% for an outlier ratio fixed to 10%. For lower outlier ratios, the precision obtained when using S1 features is slightly higher than the precision obtained when using S2 features (which will be discussed later). For higher outlier ratios, the precision decreases (more false positives are detected) but remains close to 85% for an outlier ratio equal to 20%. These results highlight the ability of the IF algorithm to provide relevant outlier scores: the parcels with the highest outlier scores are more likely to be true positives. Figure 3.21(a) shows the distribution of the detected parcels in the different anomaly categories. The majority of the detected parcels are affected by *late growth* (35%) and *heterogeneity* (25%). Anomalies coming from an error in the database (*wrong shape* and *wrong crop type* reported) are also largely detected (18.5%). To further investigate these results, Figure 3.21(b) depicts for each category the percentage of detected parcels. All parcels

of the category *wrong type* are detected, which can be understood since this anomaly strongly affects the features at the parcel-level. Using S1 features leads to detect more parcels of the category *wrong shape* when compared to using S2 features. A similar observation can be done for *vigorous crops* and *early flowering* to a lesser extent (for this outlier ratio, only few of these transient anomalies are detected).

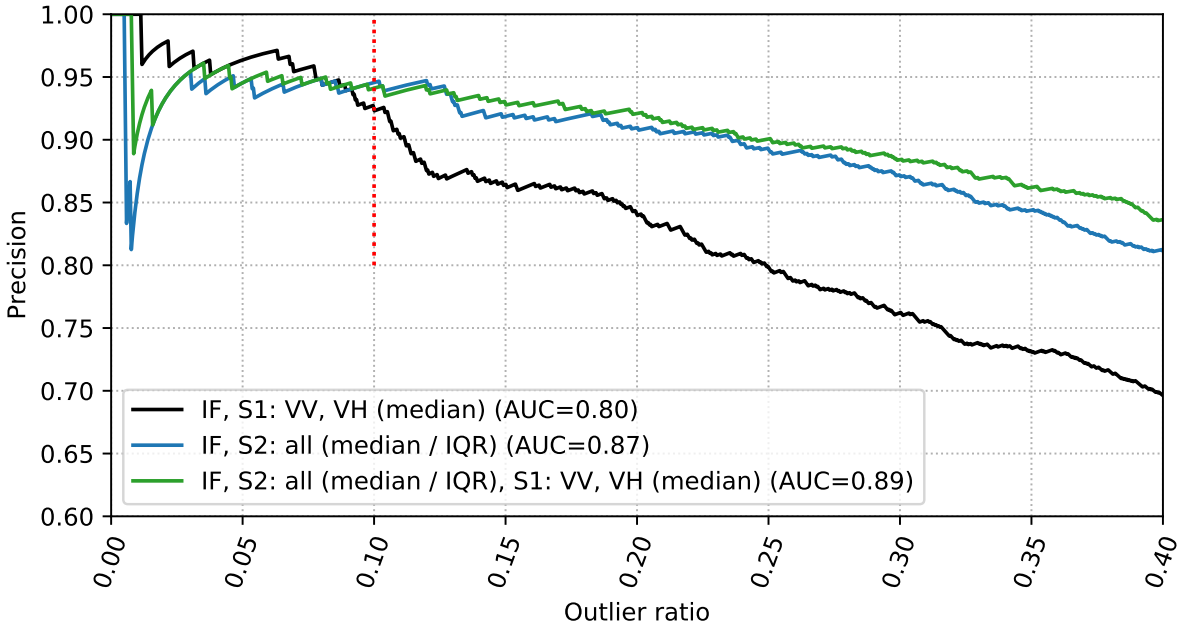
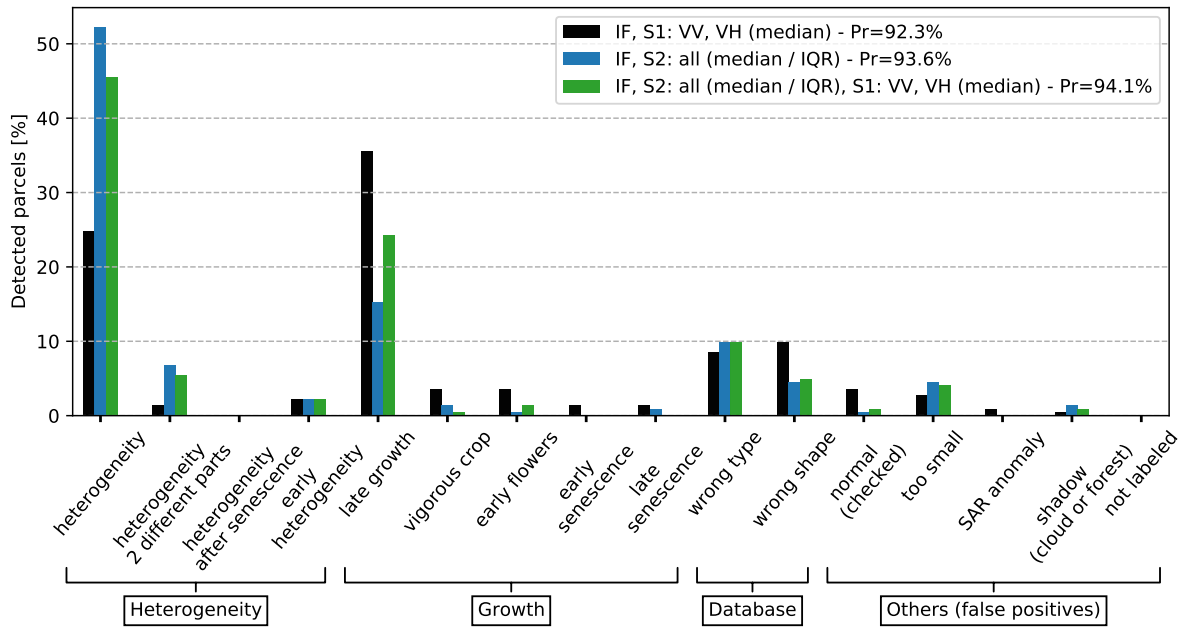
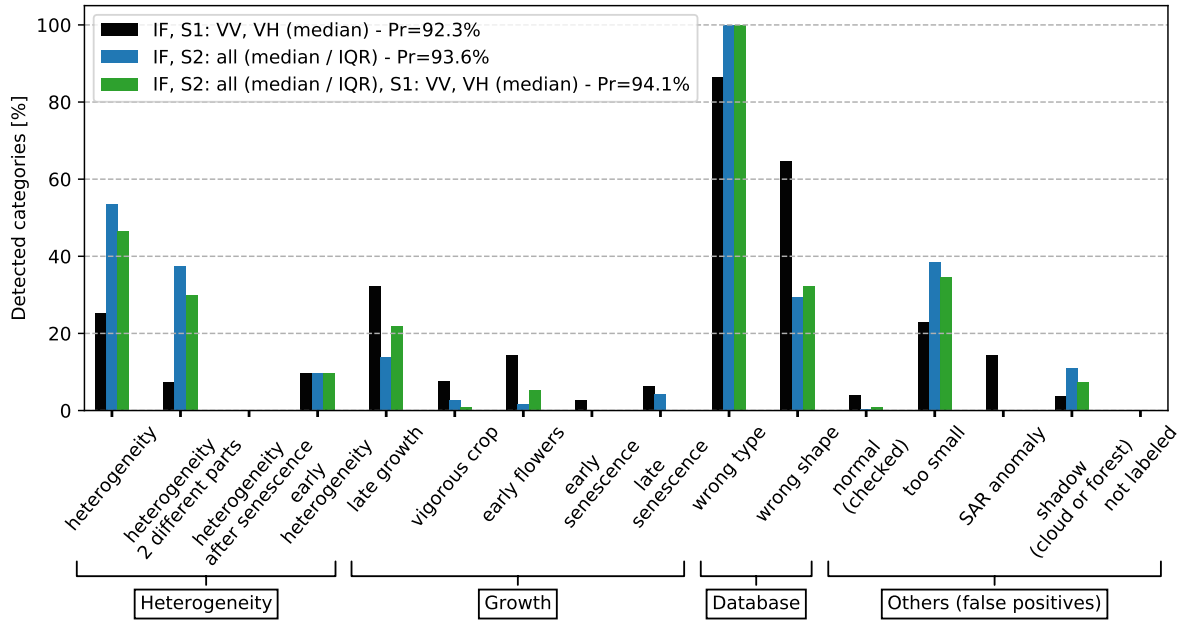


Figure 3.20: Precision vs. outlier ratio using the IF algorithm on the rapeseed parcels. Black: S1 features only, blue: S2 features only, green: S1 and S2 features jointly. The red line corresponds to the outlier ratio used in Figure 3.21

.



(a)



(b)

Figure 3.21: (a) 100×(Number of detected parcels in each category / Number of detected parcels) (b) 100×(Number of detected parcels in each category / Number of parcels in each category). The analysis is conducted using rapeseed parcels with an outlier ratio equal to 10% and the IF algorithm. Black: S1 features only, blue: S2 features only, green: S1 and S2 features jointly. The precision (Pr) of the results for each feature set is added in the legend.

3.6.1.2 Outlier detection with S2 features

Although S2 time series have lower temporal resolution when compared to S1 time series, they are useful for outlier analysis as shown in Figure 3.20 (blue curve). For an outlier ratio fixed to 20%, the precision of the detection obtained using S2 features only is still above 90%. Moreover, the average precision for outlier ratios in the range [0, 0.5] is equal to 87% whereas it is 80% when using S1 features only. For a complete growing season, having 13 S2 images is sufficient to detect a majority of relevant anomalies. However, it appears that S1 and S2 features tend to detect different types of anomalies as highlighted in Figure 3.21(a). When using S2 features, the IF algorithm detects a majority of *heterogeneous* parcels (52%) and less *late growth* parcels (15%). This observation justifies the joint use of S1 and S2 features, which is investigated below. Figure 3.21(b) shows that 40% of the parcels affected by two parts heterogeneity are detected when using S2 features (only 10% are detected when using S1 features). Moreover, a larger amount of *too small* parcels are detected when using S2 features (around 40% whereas it is close to 20% when using S1 features). This last observation should be put in perspective with the small number of parcels belonging to this category (less than 5% of the detected parcels).

3.6.1.3 Outlier detection with S1 and S2 features

One of the main objectives of this study is to investigate the joint use of S1 and S2 for outlier detection in agricultural crops. Figure 3.20 (green curve) shows that the average precision obtained when using S1 and S2 features jointly is close to 89%, which is the best performance obtained for a complete growing season analysis of the rapeseed parcels. This result means that a larger amount of relevant anomalies are detected for a given outlier ratio when compared to using S1 or S2 features separately. Moreover, it also means that the IF algorithm is able to efficiently use the characteristics of each sensor. Figure 3.21(a) shows that using S1 and S2 features jointly allows the contribution of each sensor to be accounted. In particular, late growth anomalies are more detected when compared to using S2 features only (24% vs. 15% of the detected parcels) and heterogeneous parcels are more detected when compared to using S1 features only (45% vs. 25% of the detected parcels). These observations are confirmed by Figure 3.21(b).

Overall, the best combination of features obtained throughout the study consists in using S1 and S2 features jointly. This combination exploits the strength of each sensor for crop monitoring. To be more specific, on the one hand some heterogeneous parcels are not impacting the features extracted from S1 images since this sensor is not sensitive to the color of the agricultural parcels. On the other hand, some anomalies affecting the crop growth are impacting more clearly the S1 time series that are more sensitive to the vegetation structure. Moreover, since S1 time series are dense, it is in some cases easier to detect late growth or senescence problems (*e.g.*, as mentioned for the wheat crop analysis where only few S2 images were available during the senescence phase). These results are confirmed in what follows when analyzing a different crop type.

3.6.2 Extension to wheat crops

A complementary analysis was conducted to measure the robustness of the proposed method to a change in the crop type. An experiment is presented with the selection of the best features used for rapeseed crops, *i.e.*, all the features listed in Table 3.5. The IF algorithm was used to detect abnormal wheat parcels for a complete growing season with an outlier ratio of 10%. The distribution of the detected anomalies in the different categories is depicted in Figure 3.22, which also indicates the precision obtained for each detection. Again, combining S1 and S2 data leads to the best precision (95.5%). Similar to rapeseed crops, using S1 data allows more growth anomalies to be detected when compared to S2 data only. The precision obtained using S1 features only is lower due to a higher number of SAR anomalies (i.e., 22 SAR anomalies) but the results are still accurate (precision=86.9%). As for the rapeseed analysis, no SAR anomaly is detected when using S1 and S2 data jointly. Finally, since less S2 images were available during the senescence phase, using S1 features logically leads to better detect problems affecting this growing phase and confirms the interest of using both types of features. These results confirm the interest of the proposed approach and its robustness to changes in the crop type.

Some differences were observed after analyzing the results obtained for rapeseed and wheat crops. These differences are interesting to analyze since they provide specific information for the monitoring of each crop type. For the wheat crops, the percentage of detected heterogeneous parcels is lower: when using S2 features, 31% of the detected wheat parcels belong to this category whereas 52% of heterogeneous parcels are detected for the rapeseed crops. On the other hand, the amount of detected vigorous parcels is higher (28% when using S2 features only) whereas only few vigorous parcels were detected during the rapeseed analysis. It is also interesting to note that these parcels are more easily detected using S2 features only whereas late growth anomalies are still detected in higher proportion (52%) when using S1 features only.

The fact that more late growth anomalies have been detected for wheat parcels is coherent with the observations made during the labeling, where it was noticed that late growth problems frequently have a bigger impact on the wheat parcels. A representative example is provided in Figure 3.23: the rapeseed parcel affected by a late growth anomaly has a normal vigor after the flowering phase, whereas the wheat parcel has a low vigor for the complete growing season. It was also observed that few abnormally vigorous parcels have been detected among the rapeseed dataset: this could be related to an early sowing date and a high vigor shortly after the plant emergence as pointed out in Veloso et al. (2017). Finally, the fact that few abnormally vigorous wheat parcels have been detected when using S1 features only is also coherent with the observations made in Veloso et al. (2017), where it was highlighted that the SAR signal remains stable during early growing season whereas the NDVI starts increasing after the emergence of the plant.

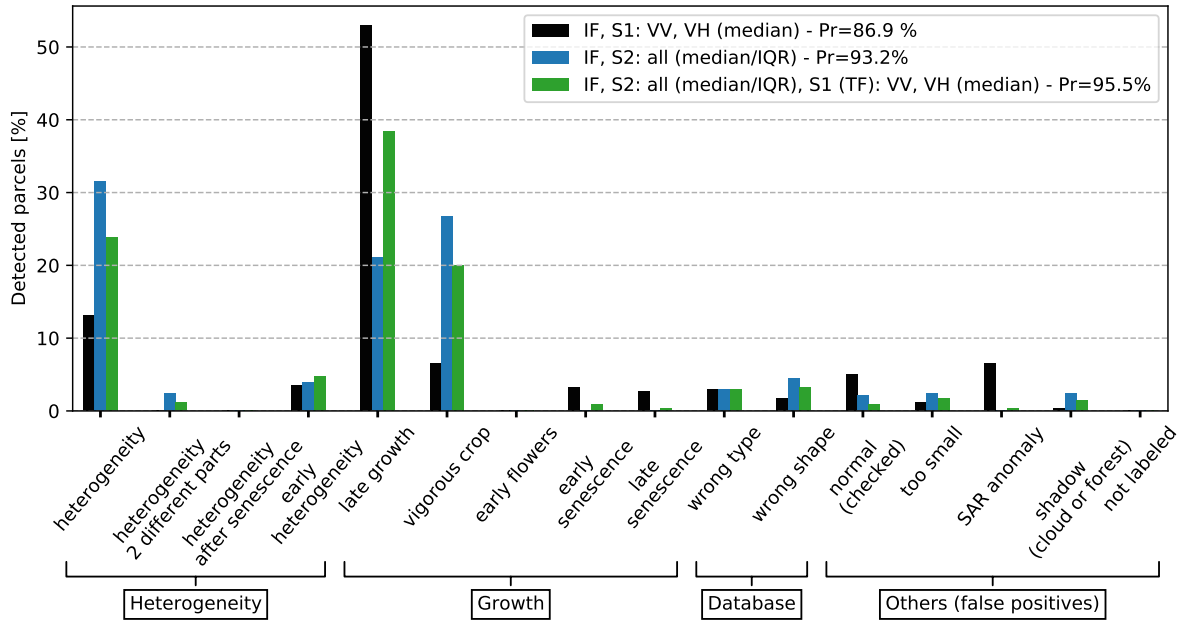
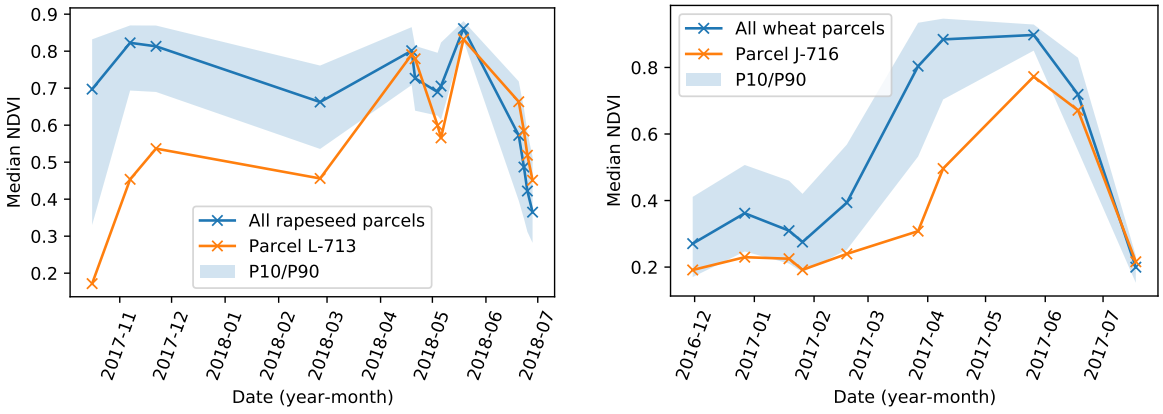


Figure 3.22: 100×(Number of detected parcels in each category / Number of detected parcels). The analysis is conducted using wheat parcels with an outlier ratio equal to 10% and the IF algorithm. Black: S1 features only, blue: S2 features only, green: S1 and S2 features jointly. The precision (Pr) of the results for each feature set is added in the legend.



(a) (b)

Figure 3.23: Median NDVI for late growth parcels (a) rapeseed parcel and (b) wheat parcel. 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 is a specific parcel analyzed.

3.7 Influence of other factors on the detection results

Various other factors that can influence the detection results were analyzed in complement of the experiments presented in the previous section. A summary of the influence of each factor is available in Table 3.6 and detailed experiments are provided in Appendix B.

Table 3.6: Summary of the influence of the different analyzed factors for the detection of anomalous crop development.

Evaluated factor	Effect and recommendation
Outlier detection algorithm	Similar results obtained with various algorithms. IF is recommended for its robustness and easy tuning.
Outlier score	Strongest anomalies have a higher outlier score than transient anomalies, which is interesting for crop monitoring.
Size of the dataset	The proposed method can work with a limited number of parcels.
Zonal statistics	Adding new zonal statistics did not improve the detection results.
Missing S2 data	The proposed method is robust to missing S2 data. Using S1 dense time series improves the results.
Mid growing season analysis	Results with high precision are obtained, early analysis is possible.
Changes in parcel delineation	Small changes in the parcel delineation do not affect the detection results

Experiments were conducted by changing the outlier ratio and analyzing their distribution among the different categories of outliers (Figure B.1). For a low outlier ratio (*e.g.*, 10%), the detected parcels are affected by strong agronomic anomalies (*e.g.*, global heterogeneity, globally low vigor). It is of crucial importance because it means that the IF algorithm attributes to these parcels the highest anomaly scores, which is relevant from an agronomic point of view. Then, parcels with lower outlier scores are affected by more transient anomalies, such as senescence problems. Using an outlier ratio equal to 20% ensures the detection of the most important anomalies among the parcels, with a low amount of false positives when using both S1 and S2 features.

An experiment was conducted by selecting randomly a subset of parcels from the original dataset in order to evaluate the robustness of the proposed methods with respect to the number of parcels in the dataset. Results are displayed in Figure 3.24, which shows the area under the Precision-Recall with respect to the number of parcels in the dataset (averaged after 50 Monte Carlo runs). In that experiment, we choose to use Precision-Recall curves instead of Precision-oultier ratio curve since the amount of parcels in the dataset is varying. It can be observed that the AUC (i.e., the average precision) obtained is stable, independently from the number of parcel in the dataset. This show that the proposed method can be used even in small dataset. We should note however that in very small dataset (e.g., 50 parcels), the

standard deviation of the AUC increase, which seems logical regarding at the small amount of potential anomalies to be detected.

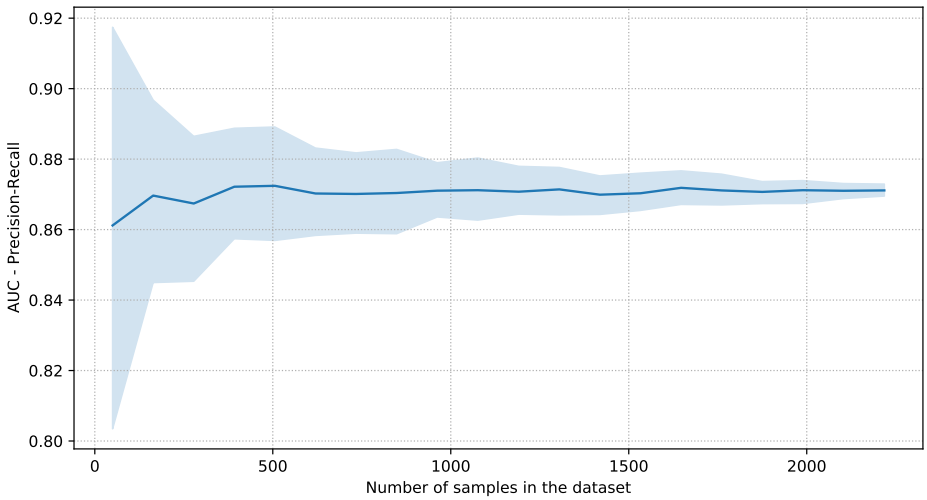


Figure 3.24: Area under the Precision-Recall curve (AUC) with respect to the number of parcels in the dataset, using the IF algorithm with S1 and S2 features. Results are averaged after 100 Monte Carlo simulations.

The impact of changing the zonal statistics used to extract parcel-level features was evaluated (Figure B.2). It appears that IQR and median statistics are of crucial importance to characterize efficiently the parcels behaviors. On the other hand, adding more subtle spatial statistics such as the kurtosis or the skewness was found to decrease the precision of the detection results.

The robustness of our method was also tested regarding the impact of missing S2 images. A good precision was obtained even with a low amount of S2 images: by using half of the S2 images, a similar precision is obtained thanks to the complementary of S1 data, which is permanently available (Figure B.4 and Figure B.5). Moreover, an outlier analysis conducted on a mid growing season (all images acquired before February) was investigated in more detail. The main reasons were to 1) measure the impact of a reduced temporal interval for the analysis and 2) investigate the interest of such analysis for early warning purposes. The results (Figure B.6 and Figure B.7) show that a large amount of abnormal parcels are detected with high precision and that the presented method can be used for an early growing season analysis.

Finally parcel delineations coming from the French Land Parcel Identification System (LPIS) were investigated to confirm the robustness of the proposed method to small changes in the parcel boundaries for the rapeseed parcels (Figure B.8 and Figure B.9). Our results confirm that the proposed method provides consistent results even when using parcel boundaries of lower precision.

3.8 Explaining the output of the IF algorithm

Interpreting the output of machine learning algorithms has become more and more important in the context of explainable artificial intelligence (XAI) (Gunning et al., 2019). XAI aims at improving the end-user experience by explaining why an algorithm arrived at a specific decision. In the context of crop monitoring and the automatic detection of anomalous crop development, using XAI can help to identify why a specific agricultural parcel was detected as abnormal. Recent improvements were made to explain the output of any machine learning algorithm using game theory leading to SHAP (SHapley Additive exPlanations) (Gunning et al., 2019). TreeExplainer, an algorithm specifically adapted to explain the output of tree based machine learning models, was proposed by Lundberg et al. (2020) and has the advantage of being fast and providing exact computation of Shapley values, which are used to explain the output of an algorithm. More precisely, Shapley values reflect the contribution of each features when attributing a score to each sample of the dataset.

An example applied to a rapeseed parcel with a late senescence is displayed in Figure 3.25 when using only NDVI statistics computed at the parcel-level. Using Shapley values attributed to each feature of a specific parcel (Figure 3.25(c)), one can easily identify the features that increase the parcel outlier score and localize in time the anomaly. In that figure, Shapley values are displayed using a color map varying from green, which corresponds to feature with SHAP values close to 0 (*i.e.*, that do not contribute to increase the abnormality of the considered sample), to red, which corresponds to features that increase the outlier score of the considered sample (i.e., here negative Shapeley values). In that case, using IQR and median of the parcel NDVI is particularly useful since these indicators have the advantage of being easily connected to the parcel behavior in terms of vigor and heterogeneity. The main limitation of this representation is that it is not possible to identify if the feature values are higher or lower than the rest of the parcels (the temporal localization is however always possible). Consequently, it is not possible to separate for instance late senescence and early senescence anomalies. Another limitation is that the features used should be easily understandable by the user, which may reduce the potential interest of using more complex features such as the one extracted from S1 data.

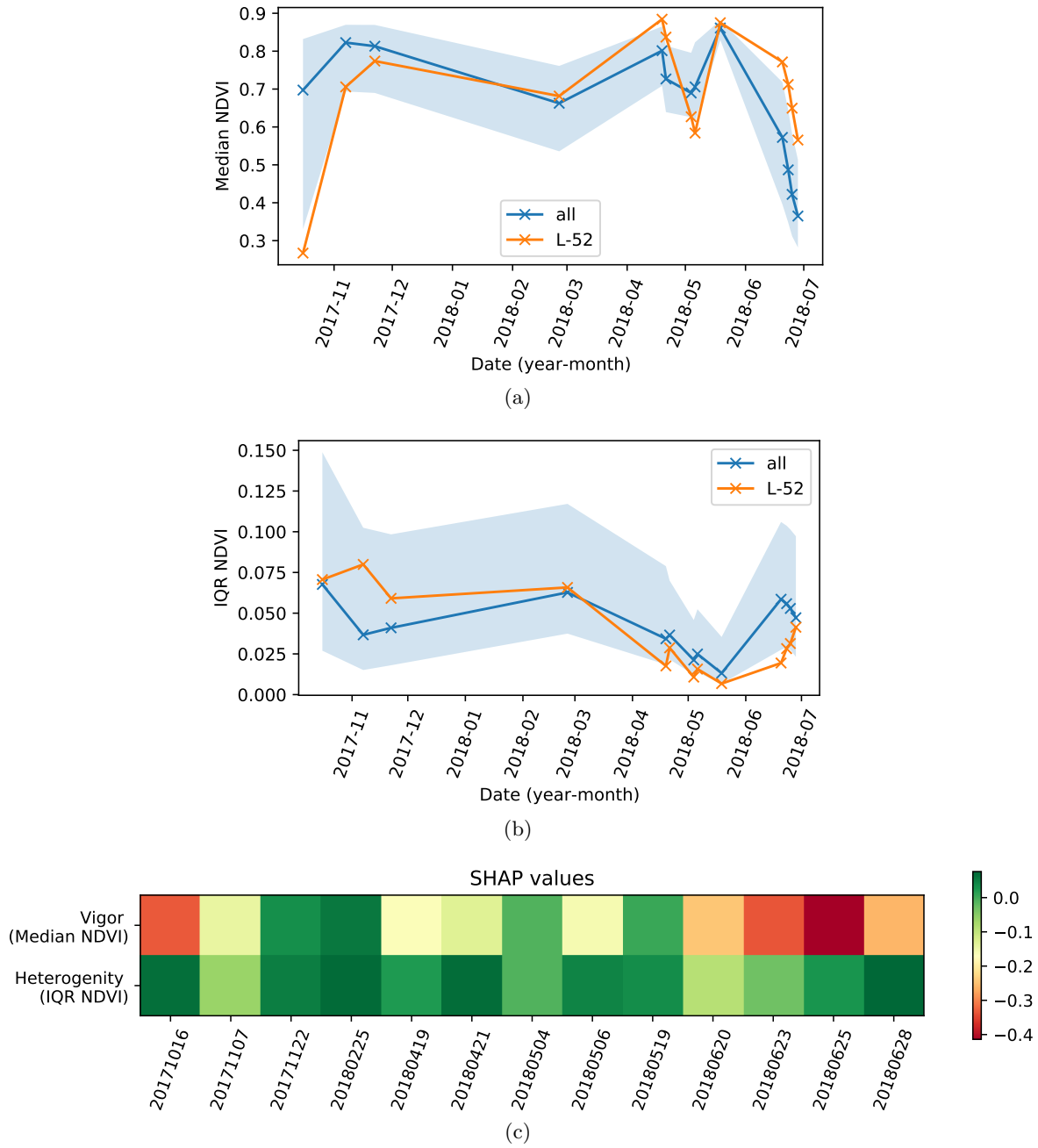


Figure 3.25: Example of a rapeseed parcel with late senescence. (a): median of the parcel NDVI (b): Interquartile range (IQR) of the parcel NDVI. The blue line is the median value of the whole dataset. The blue area is filled between the 10th and 90th percentiles. The orange line is the IQR NDVI for the analyzed parcel. (c) Shapeley computed for the parcel features, using a color map varying from green to red.

3.9 Conclusion

This chapter studied a new anomaly detection method for crop monitoring based on outlier analysis at the parcel-level using S1 and S2 images. This method is decomposed into 4 main steps: 1) preprocessing of multispectral and SAR images, 2) computation of pixel-level features, 3) computation of zonal statistics at the parcel-level for all pixel-level features at each date (these three steps are presented in Chapter 2), 4) detection of abnormal agricultural parcels using the Isolation Forest algorithm with the features extracted in step 3. The proposed method is fully unsupervised and can be used without historical data. It can be applied to different kinds of crops (such as rapeseed or wheat, considered here) and is able to detect a majority of parcels that are abnormal in an agronomic sense. Moreover, a relevant anomaly score is attributed to each parcel: agronomic anomalies affecting most of the growing season have a higher score than transient anomalies. Finally, it was shown that the proposed method can be used even with a small number of parcels, which can be valuable for operational applications which are not always conducted on large datasets.

This chapter showed that S1 and S2 features are complementary for the detection of abnormal parcels in agricultural crops. Regarding S1 features, it is recommended to use median statistics computed at the parcel-level from VV and VH backscattering coefficients. For S2 features, median and IQR statistics computed at the parcel-level from the Normalized Difference Vegetation Index (NDVI), the Green-Red Vegetation Index (GRVI), two variants of the Normalized Difference Water Index (NDWI) and a variant of the Modified Chlorophyll Absorption Ratio Index (MCARI/OSAVI) provided the best results, especially when combined with S1 features. Finally, the Isolation Forest algorithm is the outlier detection algorithm that provides the best results for identifying abnormal parcels with a simple parameter tuning.

Chapter 4

Reconstruction of Sentinel-2 Time Series with Missing Data Using Gaussian Mixture Models

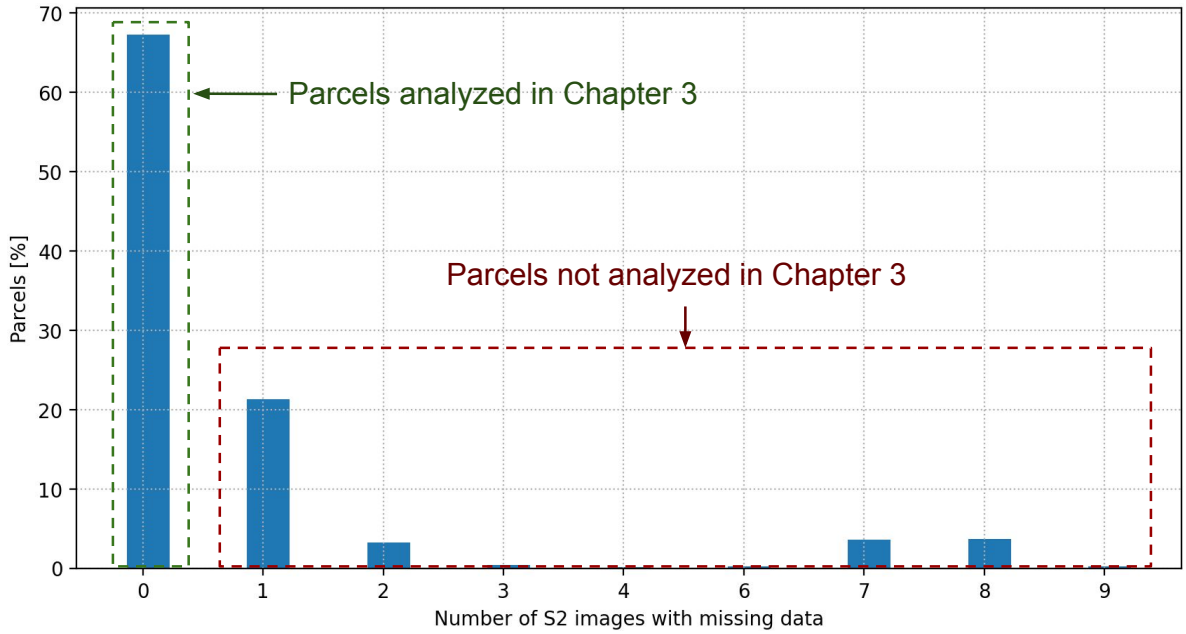
Contents

4.1 Introduction

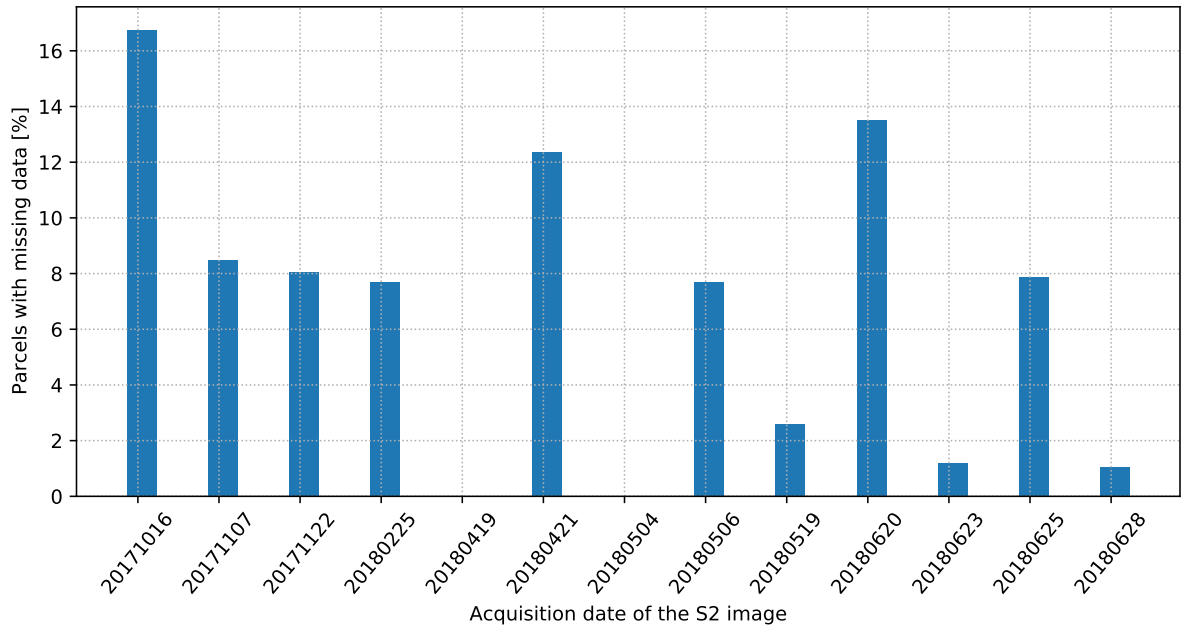
A main challenge for applications based on remote sensing is the presence of missing data. Multispectral images are particularly sensitive to this issue since they are affected by clouds (to a lesser extent, acquisition problems can also affect SAR images). As an example, Figure 4.1 illustrates the impact of missing S2 data on the rapeseed parcels analyzed in Chapter 3. Figure 4.1(a) provides the distribution of the number of S2 images with missing data among the analyzed parcels and Figure 4.1(b) shows the percentage of parcels affected by missing values for each S2 image. One can see in Figure 4.1(a) that only 67% of the parcels available for analysis have no missing data (representing 2218 parcels among the 3297 initially available), and this after selecting the least cloudy images available during the growing season. Moreover, Figure 4.1(b) shows that only 2 of the 13 S2 images selected for the analysis are not impacted by missing data problems (obviously, selecting more cloudy S2 images would lead to a higher percentage of parcels with missing data).

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). Moreover, the problem of missing data is of crucial importance when using machine learning techniques, which generally assume a complete feature matrix. It is for instance the case with the outlier detection algorithms used in Chapter 3 to detect anomalous crop developments at the parcel-level. Regarding this previous analysis, we recall here two main points related to missing data: 1) we did not select the S2 images covered by too many clouds (*i.e.*, only 10 and 13 S2 images were selected for the two growing seasons analyzed) and 2) we discarded all the parcels with missing data from the analysis, as depicted in Figure 4.1(a). Concerning the first point, we showed in Chapter 3 that it is possible to detect anomalous crop development with few S2 images. Nevertheless, exploiting the information provided by additional (but cloudy) S2 images could be interesting to capture transient events affecting the crop parcels. It could also be useful for other applications or for the end user (*e.g.*, farmers may wish to have access to consistent and timely time series for each parcel). Regarding the second point, discarding from the analysis a parcel with missing data is obviously not acceptable for operational applications, which mainly motivates the need to address the missing data problem. To that extent, this chapter focuses on the reconstruction of multiple parcel-level features extracted from S1 and S2 data (their computation is detailed in Chapter 2), when part of S2 images are missing due to the presence of clouds.

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.,



(a)



(b)

Figure 4.1: Missing data for the rapeseed analysis, (a) Distribution of the number of S2 images with missing data among the analyzed parcels (the green box correspond to the parcels analyzed in Chapter 3 whereas the parcels in the red box were discarded from this analysis due to the presence of clouds) (b) percentage of parcels with missing data for each S2 acquisition.

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 an 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. 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). 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, which is not adapter for operational services. 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. However, 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. Moreover, the Gaussian assumption used in this method can be restrictive, as will be shown in this chapter.

The method investigated in this Chapter can impute missing features computed from S2 data in a robust fashion by using Gaussian Mixture Models (GMMs). The parameters of GMMs can be estimated efficiently using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). 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 Chapter is organized as follows. Section 4.2 presents the proposed method for reconstructing missing data in features extracted from S2 images. Experimental imputation results are presented in Section 4.3. Moreover, applications to crop monitoring and the detection of anomalies in the development of rapeseed crops are presented in Section 4.4 (additional experiments were conducted on wheat crops and are available in Appendix C).

In Section 4.5, a discussion on the results and the different imputation methods is proposed. Finally, some conclusions are drawn in Section 4.6.

4.2 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. GMMs can be learned using the Expectation-Maximization (EM) algorithm, which can be naturally extended to handle missing data in a multivariate space (Dempster et al., 1977; Ghahramani and Jordan, 1994). After presenting the general principle of the EM algorithm for GMM estimation in Sections 4.2.1 and 4.2.2, Section 4.2.3 introduces 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 applications of GMMs to agricultural machine vision systems, it is worth mentioning the review proposed by Rehman et al. (2019). An interesting application of the EM algorithm to the detection of cucumber disease was also considered in Zhang et al. (2017).

4.2.1 The standard EM algorithm

Given a feature matrix $\boldsymbol{X}$ in $\mathbb{R}^{N\times p}$, where N is the number of parcels in the dataset and p is the number of features computed for each parcel, 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):

$$\log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right), \tag{4.1}$$

where $\boldsymbol{x}_n \in \mathbb{R}^p$ is a specific sample contained in the nth row of $\boldsymbol{X}$, $\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 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 $\boldsymbol{\mu}_k \in \mathbb{R}^p$, the covariance matrices $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ and the mixing coefficients $\pi_k \in]0, 1[$ of the GMM are such that $\sum_{k=1}^{K} \pi_k = 1$. The maximization of (4.1) with respect to (w.r.t.) $\boldsymbol{\theta}$ being complex, it is classical to introduce binary label vectors $\boldsymbol{z} = \{\boldsymbol{z}_1, ..., \boldsymbol{z}_N\}$ (with $\boldsymbol{z}_n \in \{0, 1\}^K, n = 1, ..., N$), 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:

$$\log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log \left[\pi_k \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right], \tag{4.2}$$

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, $\boldsymbol{\theta}^{(t)}$:

$$E[\log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{z}))|\boldsymbol{\theta}^{(t)}] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \log \left[\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right], \tag{4.3}$$

where $\gamma_{nk} = E[z_n = k|\boldsymbol{x}_n, \boldsymbol{\theta}^{(t)}]$ is referred to as the responsibility of x_n for class k. 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 $\boldsymbol{\theta}^{(t+1)}$:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} E[\log \mathcal{L}_c(\boldsymbol{\theta}; z)|\boldsymbol{\theta}^{(t)}]. \tag{4.4}$$

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 the rest of the paper.

E-step: 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:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{4.5}$$

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

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \boldsymbol{x}_n, \; \boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T, \; \pi_k^{(t+1)} = \frac{N_k}{N}, \tag{4.6}$$

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

4.2.2 Extension to handle missing data

The EM algorithm is known to be able to handle missing data since the estimation of mixture densities can be itself viewed as a missing data problem Ghahramani and Jordan (1994). 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 $m_n = m$ in the following. Using these notations, the log-likelihood of the observed

vectors can be expressed as follows:

$$\log \mathcal{L}_c(\boldsymbol{\theta}; \boldsymbol{X}^o, \boldsymbol{X}^m, \boldsymbol{z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \log \left[\pi_k \mathcal{N}(\boldsymbol{x}_n^o | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right], \tag{4.7}$$

with $\boldsymbol{X}^o$ the set of all observed variables, $\boldsymbol{X}^m$ the set of all missing variables, and $\mathcal{N}(\boldsymbol{x}_n^o | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ the marginal multivariate normal probability density of the observed sample $\boldsymbol{x}_n^o$. The E-step of the EM algorithm used for missing data computes the component responsibilities using the observed variables (Ghahramani and Jordan, 1994):

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n^o, \boldsymbol{\mu}_k^o, \boldsymbol{\Sigma}_k^{oo})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n^o, \boldsymbol{\mu}_j^o, \boldsymbol{\Sigma}_j^{oo})}. \tag{4.8}$$

In the presence of missing values, the expectation of the complete data likelihood requires the computation of additional terms due to the evaluation of $E\left[(\boldsymbol{x}_n - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k) | \boldsymbol{\theta}, \boldsymbol{x}_n^o\right]$. More precisely, the following quantities have to be computed:

$$\hat{\boldsymbol{\mu}}_{nk}^m = \boldsymbol{\mu}_k^m + \boldsymbol{\Sigma}_k^{mo}(\boldsymbol{\Sigma}_k^{oo})^{-1}(\boldsymbol{x}_n^o - \boldsymbol{\mu}_k^o), \tag{4.9}$$

$$\hat{\boldsymbol{x}}_{nk}^m = (\boldsymbol{x}_n^o, \boldsymbol{\mu}_{nk}^m), \tag{4.10}$$

$$\hat{\boldsymbol{\Sigma}}_{nk}^{mm} = \boldsymbol{\Sigma}_k^{mm} - \boldsymbol{\Sigma}_k^{mo}(\boldsymbol{\Sigma}_k^{oo})^{-1}\boldsymbol{\Sigma}_k^{mo}, \tag{4.11}$$

$$\hat{\boldsymbol{\Sigma}}_{nk} = \begin{pmatrix} \mathbf{0}^{oo} & \mathbf{0}^{om} \\ \mathbf{0}^{mo} & \hat{\boldsymbol{\Sigma}}_{nk}^{mm} \end{pmatrix}, \tag{4.12}$$

where $\mathbf{0}^{oo}, \mathbf{0}^{om}$ and $\mathbf{0}^{mo}$ are matrices of zeros of appropriate dimensions. Note that (4.9) and (4.11) are the conditional expectation and the conditional covariance matrix of the missing variables of a sample $\boldsymbol{x}_n$ assuming that it has been generated by Gaussian $\#k$, *i.e.*, $\hat{\boldsymbol{\mu}}_{nk}^m = E[\boldsymbol{x}_n^m | \boldsymbol{x}_n^o]$ and $\hat{\boldsymbol{\Sigma}}_{nk}^{mm} = \mathrm{Var}[\boldsymbol{x}_n^m | \boldsymbol{x}_n^o]$. Note also that the missing values of x_{nk} have been replaced by their expectations $\hat{\boldsymbol{\mu}}_{nk}^m$ in (4.10). Similarly, the matrix $\hat{\boldsymbol{\Sigma}}_{nk}$ of (4.12) has been filled with zeros except for the missing components, which corresponds to $\hat{\boldsymbol{\Sigma}}_{nk}^{mm}$.

In the presence of missing data, the M-step replaces the means by their imputed values and updates the covariance matrices using an additional term taking into account the missing values:

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \hat{\boldsymbol{x}}_n, \;\; \boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \left[(\hat{\boldsymbol{x}}_n - \hat{\boldsymbol{\mu}}_k)(\hat{\boldsymbol{x}}_n - \hat{\boldsymbol{\mu}}_k)^T + \hat{\boldsymbol{\Sigma}}_{nk}\right], \;\; \pi_k^{(t+1)} = \frac{N_k}{N}. \tag{4.13}$$

More details about the EM algorithm for the GMM with missing data can be found for instance in Ghahramani and Jordan (1994). It is also 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.

4.2.3 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 associating 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, which prevents its use for crop monitoring: 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 outliers (Liu et al., 2012). As shown in Chapter 3, this algorithm was found to be efficient to detect relevant abnormal parcels 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)

$$\boldsymbol{\mu}_j^{t+1} = \frac{\sum_{n=1}^{N} w_n \gamma_{nk} \hat{x}_{nk}}{\sum_{n=1}^{N} w_n \gamma_{nk}}, \boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{N} w_n^2 \gamma_{nk} \left[(\hat{x}_n - \hat{\boldsymbol{\mu}}_k)(\hat{x}_n - \hat{\boldsymbol{\mu}}_k)^T + \hat{\boldsymbol{\Sigma}}_{nk} \right]}{\sum_{n=1}^{N} w_n^2 \gamma_{nk}}, \pi_k^{(t+1)} = \frac{N_k}{N}. \tag{4.14}$$

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

$$w_n = \frac{1}{1 + \exp\left(\alpha(\text{score}_{\text{IF}}(\hat{\boldsymbol{x}}_n) - \text{th})\right)}, \tag{4.15}$$

where α and th are two constants to be fixed by the user. Motivations for using the sigmoid (4.15) 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 (4.15) 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 the anomaly score is depicted in Figure 4.2 for $\alpha = 50$ and $\text{th} = 0.5$.

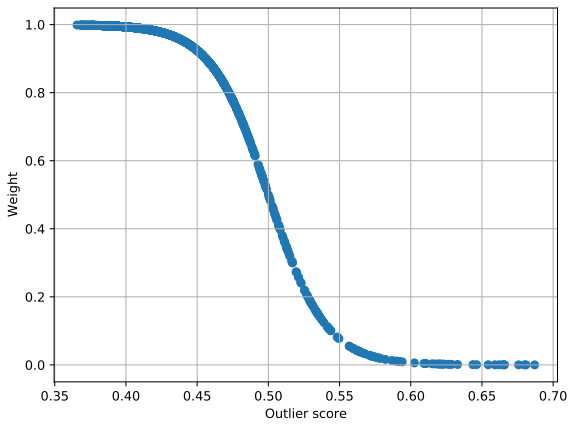
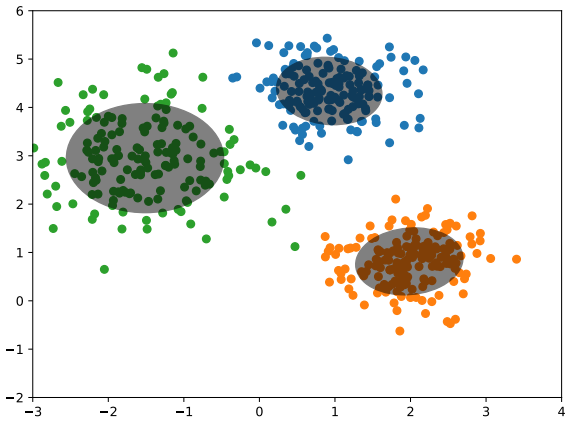


Figure 4.2: Variation of the weight w_n versus the outlier score attributed by the IF algorithm, with $\alpha = 50$ and th $= 0.5$.

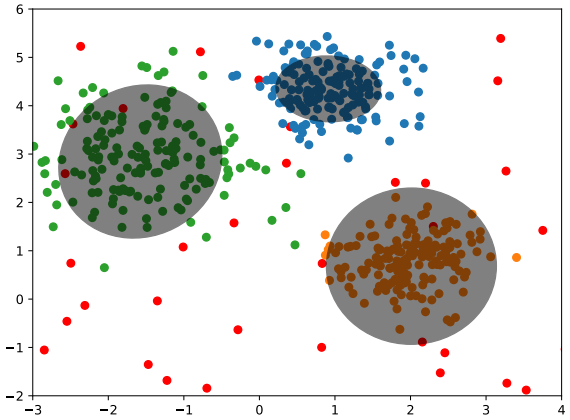
To illustrate the interest of a robust GMM estimation, a toy example is provided in Figure 4.3. The initial dataset presented in Figure 4.3(a) is contaminated by outliers (in red) in Figure 4.3(b). One can see that the GMM estimation is influenced by the presence of outliers, with an overestimation of the cluster covariances (in particular for the orange cluster). In Figure 4.3(c), we used the proposed robust GMM. One can observe that in that case the estimation of the parameters is not impacted by the outliers.

4.2.4 Regularization of the covariance matrices

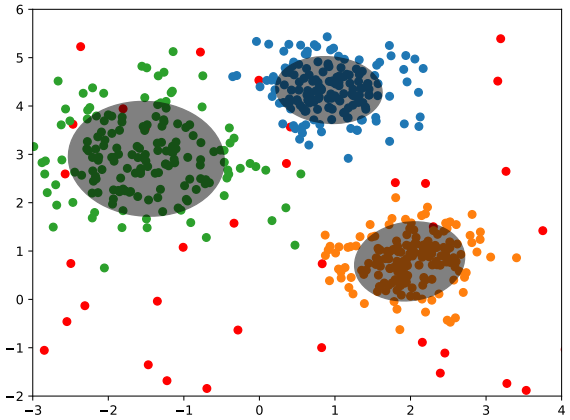
Learning the parameters of a GMM can be subject to instabilities, especially when the covariance matrices are ill-conditioned (in some extreme cases, the covariance matrix cannot even be inverted). A heuristic strategy for regularizing a covariance matrix consists in adding a small constant to its diagonal 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 $[a_{ij}bQ_id_i]$ (see Bouveyron et al. (2007) for details). The idea behind this model is to use eigendecompositions of the covariance matrices $\mathbf{\Sigma}_k = Q_k\Delta_kQ_k^T$ 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)



(b)



(c)

Figure 4.3: Toy example with 3 clusters: (a) GMM estimation without outliers, (b) GMM estimation in the presence of outliers (red points) and (c) Robust GMM estimation in the presence of outliers.

4.3 Imputation results

This section compares both robust and non-robust GMM imputation methods with imputations 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 4.2.4. Various other imputation methods (gap filling, autoencoders, multiple imputations, soft imputation) were tested and are discussed in Section 4.5. The results presented here focus on the imputation of multispectral S2 time series. However, the same method could be used to reconstruct S1 features as well. Finally, note that each feature was scaled in the range $[0, 1]$ (before performing GMM and KNN imputations). Features in natural scale can then be of course retrieved by using the inverse transformation.

4.3.1 Simulation scenarios

In order to evaluate the performance of missing data reconstruction with a controlled ground truth, we have removed some existing features in the dataset introduced in Chapter 2. 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 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 section.

Table 4.1: Summary of the experiments conducted in this chapter in terms of percentage of cloudy S2 images and 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
4.3.3	Convergence of the EM algorithm	8% (1 S2 image)	50%
4.3.4	Effect of the percentage of S2 images affected by missing values	Varies in [0, 70]%	50%
4.3.5	Effect of adding irrelevant samples	23% (3 S2 images)	50%

4.3.2 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:

$$\text{MAE} = \frac{\sum_{i=1}^{N_m} |f_i - \hat{f}_i|}{N_m}, \tag{4.16}$$

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).

4.3.3 Parameter tuning and convergence

The hyperparameters used for the different reconstruction algorithms are reported in Table 4.2, with more details included what follows.

Table 4.2: 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

4.3.3.1 GMM imputation

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 avoid to manually choose 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 from these Monte Carlo runs (evaluated using all the reconstructed features for the parcels with missing data) is displayed in Figure 4.4. This figure shows that the values of MAE are very similar, varying in the interval [0.02178, 0.02186], indicating that the EM algorithm is not very sensitive to its initialization for the reconstruction of vegetation indices (VI) at the parcel level.

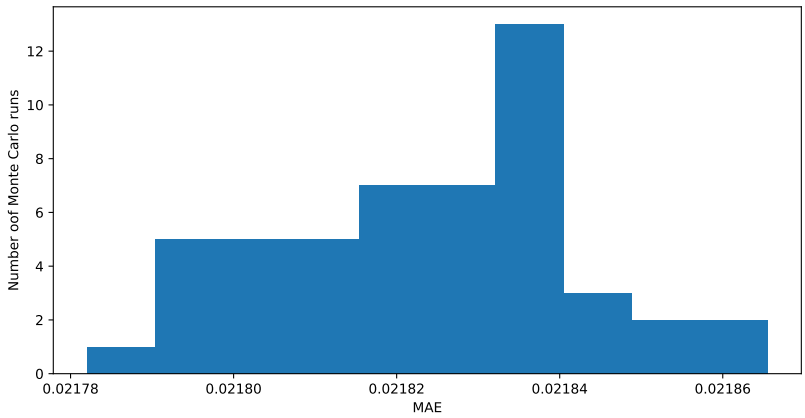


Figure 4.4: Histogram of MAE obtained after 50 Monte Carlo runs (with different initializations) on the same dataset.

4.3.3.2 KNN imputation

The KNN imputation method (Troyanskaya et al., 2001) 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).

4.3.4 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 Chapter 3. 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 4.5. 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 4.5(a) whereas Figure 4.5(b) and (c) show specifically the MAE of the median and IQR NDVI. Note that in Figure 4.5(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. Our conclusions are summarized below:

- One can observe that the GMM imputation outperforms the KNN imputation, with accurate 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, which makes sense since the dataset contains strong outliers (*e.g.*, error in the crop type reported).
- Looking specifically at the median NDVI (Figure 4.5(b)), it appears that using S1 data is particularly useful, especially when there is a high amount of missing S2 features.
- The reconstruction of IQR statistics is not favored by the use of S1 data, as shown in Figure 4.5(c). For this statistics, the robust GMM provides a lower MAE than the classical GMM.

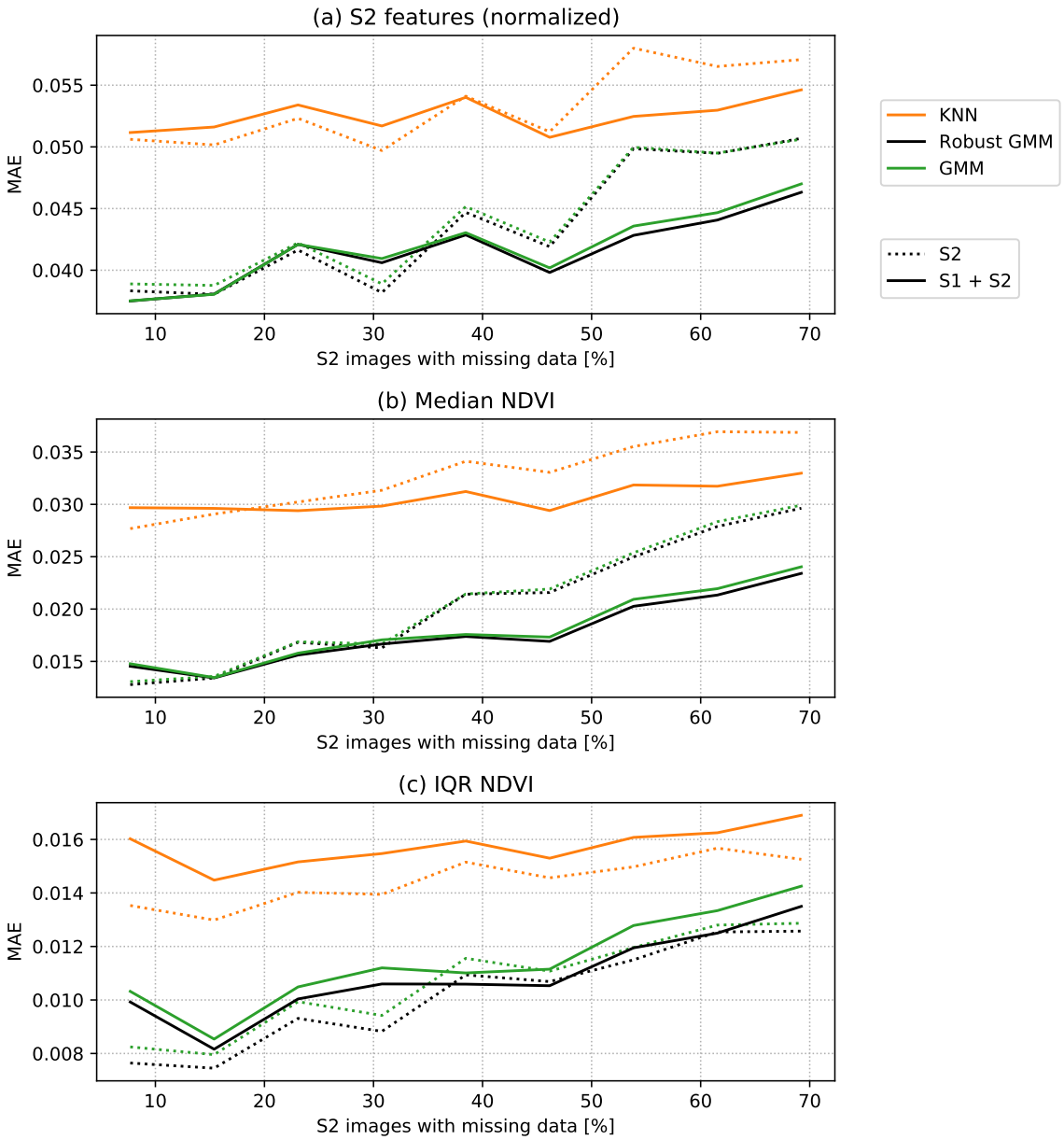


Figure 4.5: MAE for rapeseed vegetation indices versus the percentage of missing images. X-axis: percentage of S2 images with missing values. Y-axis: MAE for (a) the normalized S2 features (all the S2 indicators are considered), (b) the median of NDVI and (c) the IQR of the NDVI (computed at the parcel 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.

4.3.5 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), which is available in open license.

As an example, an imputation has been conducted on the rapeseed dataset by adding 5% of contaminated samples (coming from wheat, maize and barley crops). For this experiment, 3 S2 images have missing data affecting 50% of the crop parcels. To illustrate how the robust GMM operates, Figure 4.6 provides the distribution of the outlier scores given by the IF algorithm to each sample, and their associated weight within the robust GMM. One can see that all the parcels coming from a non-rapeseed crop type have a high outlier score (i.e., generally above 0.5), implying a reduced weight when learning the GMM. Some specific rapeseed parcels have also a high outlier score and consequently a reduced weight. These parcels correspond to strong anomalies in the crop developments or errors in the crop type reported, as studied in Chapter 3.

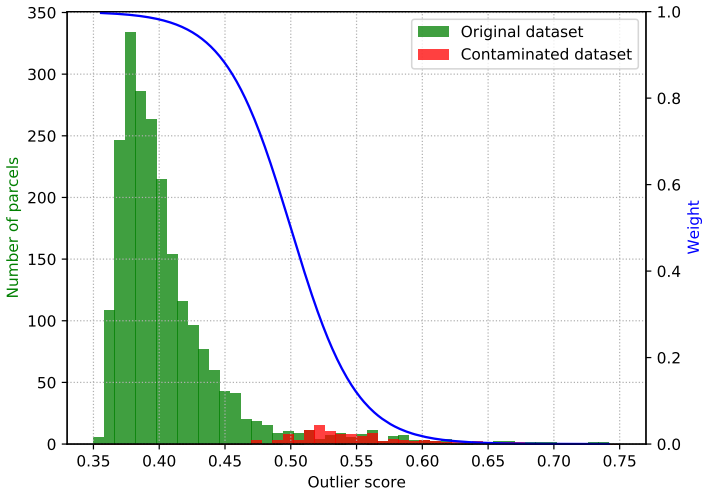


Figure 4.6: Distribution of the outlier scores given by the IF algorithm within the Robust GMM imputation. Parcels coming from the rapeseed dataset are displayed in green, whereas parcels coming from a different crop type are displayed in red. The weight attributed by the Robust GMM to each sample with respect to their outlier score is superposed in blue. For this experiment, 3 S2 images have missing data affecting 50% of the crop parcels.

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 4.7, 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. 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 (this cannot be observed in Figure 4.7, which shows MAEs averaged over the whole dataset). Consequently, using the robust GMM imputation is recommended in practice, especially if the dataset contains some irrelevant samples.

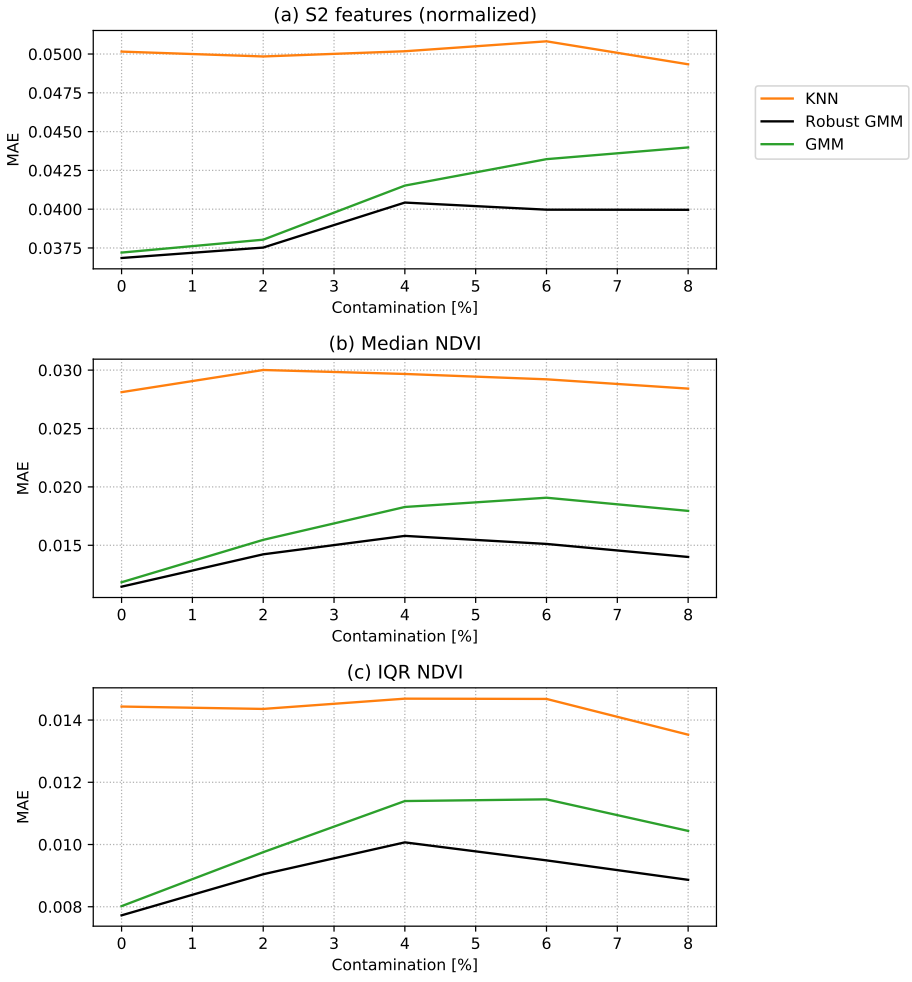


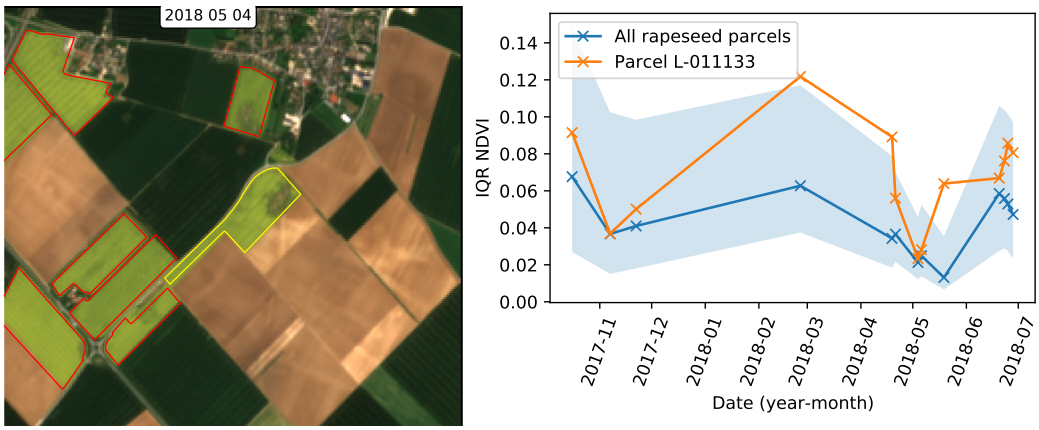
Figure 4.7: Median of MAE versus the percentage of contamination in the dataset (i.e., coming from non-rapeseed crops) after 50 MC runs for (a) the normalized S2 features (all the S2 indicators are considered), (b) the median of NDVI and (c) the IQR of NDVI (computed at the parcel level). Results are 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.

4.4 Application to crop monitoring

This section evaluates the interest of the proposed imputation method for crop monitoring. In Section 4.4.1, the detection of anomalous crop development is conducted in the presence of missing data. This addresses a main issue of the method proposed in Chapter 3, which use outlier detection algorithm that do not handle missing values in the feature matrix. In Section 4.4.2, we analyze new rapeseed parcels with missing data that were previously discarded from the analysis conducted in Chapter 3. Finally, in Figure C.7, the interest of the proposed imputation method to increase the temporal resolution of the S2 features is evaluated.

4.4.1 Detection of anomalous crop development in the presence of missing data

In Chapter 3, the rapeseed parcels have been analyzed to detect potential anomalies in their development. An example of a heterogeneous parcel is depicted in Figure 4.8 (a further analysis showed that a part of the parcel was damaged during winter). Each parcel has been labeled by an agronomic expert as true positive (relevant anomaly to be detected) or false positive (not relevant for crop monitoring). Parcels with abnormal behavior are detected using the IF algorithm, which computes anomaly scores using the feature matrix (whose construction is detailed in Chapter 2). Since the IF algorithm cannot be used if the feature matrix has missing values, we propose to impute missing data to address this issue.



(a) (b)

Figure 4.8: (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.

The following experiment 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 4.9 (more details on this metric were provided in Section 4.3.2). 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, since there are few strong outliers). It is interesting to note that discarding S2 images affected by missing values reduces the detection performance. This shows that imputation methods are able to reconstruct the VI with sufficient accuracy to detect the abnormal crop parcels whereas important information on the parcel behavior seems to be lost without the reconstruction step.

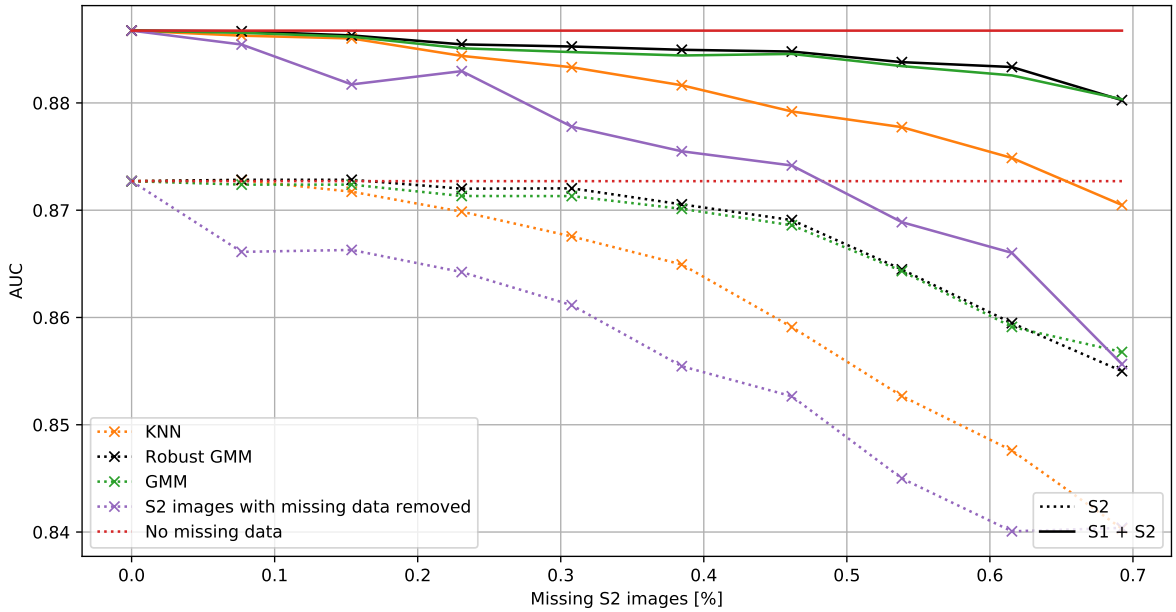


Figure 4.9: Area under the precision vs. outlier ratio curve (AUC) w.r.t. the percentage of cloudy S2 images (50% of the parcels in a cloudy S2 image have missing data, i.e., do not contain S2 features). Results in dotted lines are obtained using S2 features only whereas solid lines correspond to the joint use of S1 and S2 data. All results are averaged using 50 MC runs.

4.4.2 Analyzing new crop parcels previously removed from the database

Using the proposed imputation strategy, it is now possible to analyze the rapeseed parcels with missing data that were discarded from the analysis conducted in Chapter 3 (see the introduction of this Section and Figure 4.1). In the following experiments, a total of 3297 parcels have been analyzed (1079 parcels have at least one S2 image with missing data and 2218 have no missing data). Table 4.3 shows the precisions obtained using the anomaly detection strategy of Chapter 3 with outlier ratios equal to 10 and 20% and features reconstructed with the robust GMM or KNN methods (S1 and S2 features are used jointly). The parcels never detected before were labeled following the method described in Chapter 3. In most of the cases, the available features were sufficient to label with confidence these parcels. Overall, a high precision is observed with both KNN and GMM imputations, confirming results of Section 4.4.1. Thus, we can conclude that the parcels previously discarded can be analyzed with accuracy thanks to the imputation of the feature matrix. Note that the new parcels detected as abnormal and labeled as false positives are rare (14 parcels in total). These parcels were difficult to label with confidence because of the lack of data (without ground truth the imputed values cannot be validated).

Table 4.3: Precision (Pr.) of the detection results obtained using the IF algorithm with S1 and S2 features and outlier ratios equal to 10 and 20%. The third column indicates the precision for the parcels never analyzed before (with their number into parentheses).

Imputation	Outlier ratio	Pr. - all	Pr - samples with missing data (#)
KNN	10%	96.4	99.1 (116)
Robust GMM	10%	96.9	99.2 (128)
KNN	20%	94.2	96.1 (231)
Robust GMM	20%	94.2	96.0 (241)

A concrete example of the imputation obtained for a rapeseed parcel with missing data is provided in Figure 4.10. For this parcel, only 5 S2 images can be exploited. Even with this very limited number of images, an obvious late growth is visible when looking at the remaining S2 images (illustrated in Figure 4.11) and the corresponding S2 features (Figure 4.10(a,b)). This growth problem is also visible using the S1 times series (especially the VH backscattering, Figure 4.10(d)). Without ground-truth, it is difficult to know which imputation method captures better the actual behavior of the crop parcel. However, the GMM imputation seems more coherent when looking at the S1 times series and provides smoother imputed time series (especially for the period between May and July, where more data are available for comparison).

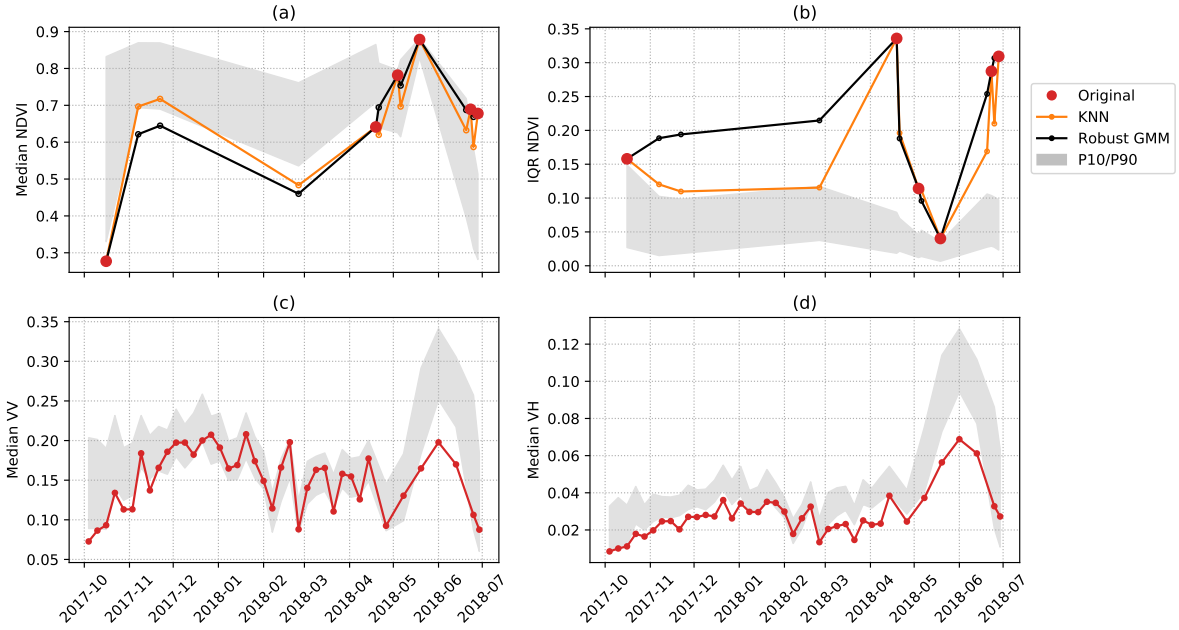


Figure 4.10: For a specific rapeseed parcel, imputation of (a) median NDVI, (b) IQR NDVI, and time series of (c) median VV (S1) and (d) median VH (S1). Actual values are plotted in red dots. The gray area is filled between the 10th and 90th percentiles values of the whole dataset.

Figure 4.11: A rapeseed parcel (yellow boundaries) affected by growth problems analyzed in Figure 4.10 (image acquired in April 2018).

4.4.2.1 Increasing the temporal resolution of the S2 features

So far, only 13 S2 images have been used to analyze the growing season of the rapeseed parcels. With reliable imputation methods, adding more cloudy S2 images seems legitimate to fully exploit the information available on the crop parcels. Increasing the temporal resolution of the S2 features can be beneficial for post-analysis and crop monitoring in general (i.e., for farmers and stakeholders who need timely information about the parcels and are not only interested in the detection of the most anomalous parcels). In what follows, two specific examples are provided to illustrate the interest of adding new S2 images. In total, 8 new cloudy S2 images were added to the database, for a total of 21 S2 images. Note that complementary results showing that adding these new images do not impact the detection of anomalous crop development are provided in Appendix C.

A first example illustrating the interest of using additional S2 images is displayed in Figure 4.12 (the parcel analyzed is the same than the parcel analyzed in Figure 4.10). One can appreciate with more details the crop development within the whole growing season, especially during winter. The additional ground-truth acquired at the beginning of the season confirms that this parcel has growth problems.

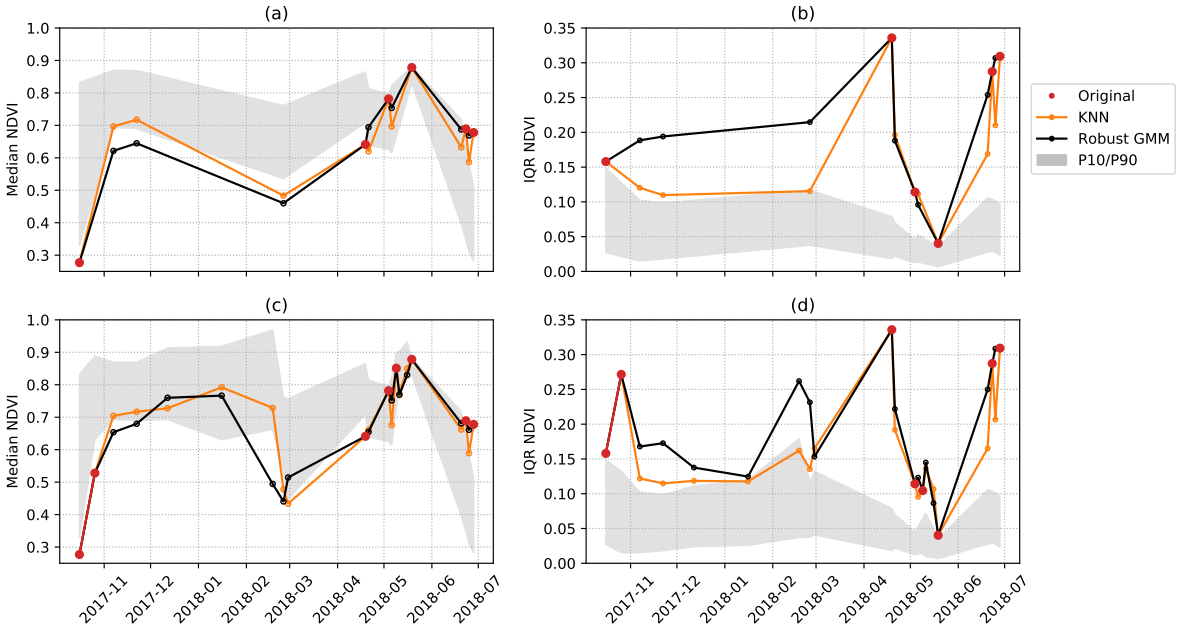


Figure 4.12: For a specific rapeseed parcel, imputation of (a,c) median NDVI, (b,d) IQR. For (a,b), only 13 S2 images are used whereas 21 images are considered for (c,d). The gray area is filled between the 10th and 90th percentiles values of the whole dataset.

Another example (less extreme since more ground-truth is available) is displayed in Figure 4.13. One can observe the interest of data imputation with an increased amount of S2 data to appreciate more accurately the different phases of the growing season. Identifying these phases is particularly difficult for crops such as rapeseed, due to fast and abrupt changes in the crop phenology.

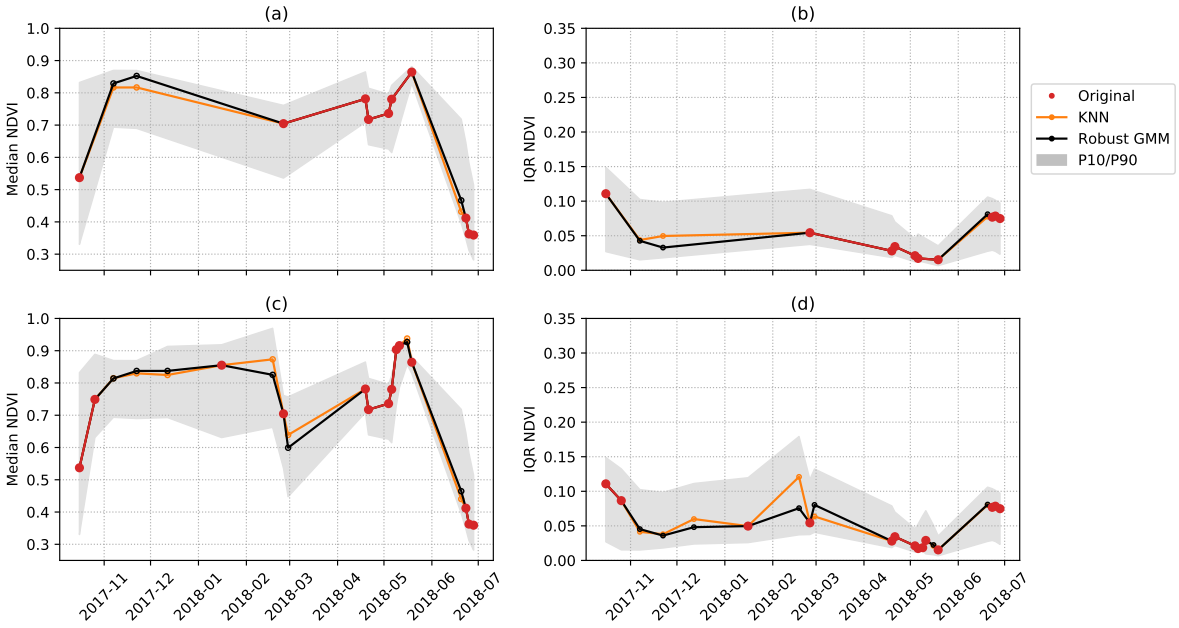


Figure 4.13: For a specific rapeseed parcel, imputation of (a,c) median NDVI, (b,d) IQR. For (a,b), only 13 S2 images are used whereas 21 images are considered for (c,d). The gray area is filled between the 10th and 90th percentiles values of the whole dataset.

4.5 Discussion

This section provides some comments about the results obtained in this Chapter. Additional experiments (available in Appendix C) are also discussed.

4.5.1 Analysis of the presented results

The experiments conducted in this study show that 1) GMM imputations outperform the KNN method and 2) using a robust GMM is of crucial importance in the presence of strong outliers. Thus, 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 MAEs varying from 0.035 to 0.042 (depending on the region analyzed) were 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 here. Moreover, some of these studies focus on the regression of NDVI time series using SAR data (Garioud et al., 2021). Taking these differences into account, obtaining an MAE close to 0.013 (resp.

to 0.019) when imputing the median NDVI for rapeseed (resp. wheat) crops is nevertheless encouraging (see Table C.1 and Table C.2 in Appendix C).

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 Appendix C for rapeseed and wheat parcels (Figure C.1 and Figure C.2). The heterogeneity of a parcel (summarized using IQR at the parcel-level) is less linked to S1 data, which confirms previous results obtained in Chapter 3. 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.

When removing features from one of the S2 image (Figure C.3 and Figure C.4), 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 high 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 2017 can be explained by the beginning of the senescence, which leads to abrupt changes in the crop behavior.

4.5.2 Other imputation methods

Other strategies for the imputation of missing data were tested without bringing any improvement compared to the proposed method (see Figure C.5). 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 chapter. These poor results are 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 Scikit-Learn Python library (Pedregosa et al., 2011). Similarly to the KNN imputation, MICE provides reconstruction results significantly less accurate than those obtained using the proposed GMM imputation. Deep learning methods were also tested without success due to the small number of parcels in the dataset (in particular, we considered a classical structure referred to as denoising autoencoders and studied in Vincent et al. (2008); Pereira et al. (2020)).

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.

4.5.3 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 our experiments 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).

4.6 Conclusion

This chapter 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 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. It was also shown 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 Chapter 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 chapter confirmed that using additional Sentinel-1 (S1) features can improve imputation results, especially to reconstruct S2 features at the parcel level, such as the Normalized Difference Vegetation Index (NDVI). 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 prevents to detect many parcels with abnormal crop development. Moreover, it was shown that the proposed imputation method can be used to increase the temporal resolution of the S2 features, which is important for crop monitoring in general.

An interesting perspective could be 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). Finally, using the proposed imputation method with dense S2 time series was found interesting to increase the temporal resolution of the S2 features, which is always valuable for crop monitoring in general. Combining the proposed method with smoothing or gap-filling techniques could be another relevant perspective and could be for instance useful to reduce the problems of undetected clouds.

Chapter 5

Towards Temporal Approaches for the Detection and Localization of Anomalous Crop Development

Part of this chapter has been adapted from the journal paper León-López et al. (2021).

Contents

5.1 Introduction

In Chapter 3, a strategy has been proposed to detect crop parcels with anomalous phenological development. This detection is made using the IF algorithm, which attributes to each parcel an outlier score proportional to its degree of abnormality. Although this strategy has proven to be relevant, it seems legitimate to ask whether taking into account the temporal structure of the data could be beneficial to this analysis. Indeed, the IF algorithm does not make any assumption regarding the temporal relationship between the different features and only attributes a unique outlier score for each parcel. On the other hand, using temporal approaches could provide valuable information, *e.g.*, to locate in time the anomalies. In this Chapter, we explore the interest of using such approaches for the detection and localization of anomalous crop development. In particular, we propose to use an ensemble of Hidden Markov models (HMMs) for this task, taking advantage of their ability to efficiently model dynamic phenomena. The presented method was first investigated in León-López et al. (2021) (collaboration made during this thesis) and was slightly adapted to our use cases, in particular to be fully unsupervised.

5.2 Hidden Markov Models for the Detection and Localization of Anomalous Vegetation Development

HMMs (Baum and Petrie, 1966) have been largely used to analyze time series, *e.g.*, for speech recognition (Rabiner, 1989). Thanks to their ability to model dynamic processes, they have also been used in remote sensing in the context of multi-temporal analysis, as for instance for the classification of crops (Siachalou et al., 2015) or to model vegetation dynamics (Viovy and Saint, 1994). In what follows, we first introduce the main idea behind HMMs. In a second step, we propose an adaptation of this idea to detect and localize anomalies.

5.2.1 HMM learning

An HMM is a statistical model in which the modeled system is assumed to be a Markov process with unknown parameters. Lets call $\boldsymbol{Y}$ the modeled Markov process, with unobservable states (i.e., "hidden states") and $\boldsymbol{X}$ another process whose behavior depends on $\boldsymbol{Y}$ (Rabiner, 1989). The goal of HMM is to model $\boldsymbol{Y}$ by observing $\boldsymbol{X}$. Let $\mathcal{S} = \{s_1, s_2, ..., s_d\}$ be the D states of the model and T the length of the observed time series (theoretically, T can vary for each observed sequence). Formally, an HMM can be described by the unknown parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$, where $\boldsymbol{\pi} \in \mathbb{R}^D$ is the initial probability vector defining the initial probabilities of the system to be in the different states, $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ is the transition probability matrix defining the probabilities of the hidden latent variables to change from one state to another, and $\boldsymbol{B} \in \mathbb{R}^{D \times T}$ is the emission probability matrix, which provides the probability of observing a given value in state s (León-López et al., 2021).

As an example, let $\boldsymbol{X}^{(n)}$ be the times series of parcel n ($\boldsymbol{X}^{(n)}$ can be multivariate, *e.g.*, considering median NDVI and IQR NDVI). The sequence of hidden states of the parcel n across time is denoted $\boldsymbol{Z}^{(n)} = \{z_1^{(n)}, z_2^{(n)}, ..., z_T^{(n)}\}$, with each $z_t^{(n)} \in \mathcal{S}$. For brevity, we will omit the subscript $^{(n)}$ and use the notations $z_t^{(n)} = z_t$ and $x_t^{(n)} = x_t$ in the following. The elements of the transition probability matrix $\boldsymbol{A}$ are defined as $a_{ij} = \mathrm{P}(z_t = s_i | z_{t-1} = s_j)$, which corresponds to the probability transition from state s_i to state s_j, with $i, j \in \{1, ..., D\}$. Furthermore, the emission probability density $b_i(x_t)$ denotes the probability density function of x_t given that x_t is in the state s_i. Here, the emission probability densities are assumed to be mixtures of K multivariate normal densities. The likelihood of a given parcel is then defined as:

$$\begin{aligned} \mathrm{P}(\boldsymbol{X}^{(n)}|\boldsymbol{\theta}) &= \sum_{\text{all } \boldsymbol{Z}^{(n)}} \mathrm{P}(\boldsymbol{X}^{(n)}|\boldsymbol{Z}^{(n)}, \theta)\mathrm{P}(\boldsymbol{Z}^{(n)}, \theta) \\ &= \sum_{\text{all } \boldsymbol{Z}^{(n)}} \pi_{z_1} b_{z_1}(x_1) a_{z_1,z_2} ... a_{z_{T-1},z_T} b_{z_T}(x_T). \end{aligned} \tag{5.1}$$

As explained in Rabiner (1989), one can interpret the above equation as follows. At time $t = 1$, we are in state z_1 with probability π_{z_1} and generate observation x_1 with probability $b_{z_1}(x_1)$. At time $t = 2$, a transition from state z_1 to z_2 occurs with probability $a_{z_1 z_2}$ and generate observation x_2 with probability $b_{z_2}(x_2)$. This process continues until the generation of observation x_T, with probability $b_{z_T}(x_T)$. The calculation of (5.1) being too computationally intensive, an efficient procedure known as the forward-backward procedure (Baum and Eagon, 1967) is used in practice to determine $\mathrm{P}(\boldsymbol{X}^{(n)}|\boldsymbol{\theta})$.

Finally, the HMM parameter vector $\boldsymbol{\theta}$ is estimated by maximizing the log-likelihood when considering all the training samples (here, the N crop parcels contained in the learning database):

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log \sum_{n=1}^{N} \mathrm{P}(\boldsymbol{X}^{(n)}|\boldsymbol{\theta}). \tag{5.2}$$

This maximization is conducted with a special case of the EM algorithm, which makes use of the forward-backward procedure known as the Baum-Welch algorithm (Rabiner, 1989).

5.2.2 Anomaly detection based on an ensemble of HMMs

In what follows, we detail a HMM-based method for the detection and localization of anomalies in time series (León-López et al., 2021). First, we introduce the general idea of the proposed method, which is based on learning an ensemble of HMMs. Second, we provide a general detection strategy at the parcel-level for a complete growing season analysis. Finally, we show that it is also possible to analyze sub-sequences of the analyzed time series to localize the anomalies in time, which is the main interest of the proposed approach.

1. **HMM ensemble learning:** the main originality of the proposed method is to build an ensemble of L HMMs, which are estimated on subsets of the training dataset. The number of samples in each subset is denoted N_s. Building various HMM models allows us to consider different possible underlying structures in the data to better explain them. This idea is for instance used in the famous classification method Random Forest (Learning, 2001) or in the IF algorithm and have been previously used with HMMs, *e.g.*, for clustering (Hamdi and Frigui, 2015). In León-López et al. (2021), a strong assumption was made on the availability of training data composed of normal parcels to learn the ensemble of HMMs. Here, we propose to build the training dataset using the IF algorithm by selecting the parcels with the lower IF scores. This allows us to have a fully unsupervised method without a need for a manual selection of training samples. This learning procedure is summarized in Figure 5.1.

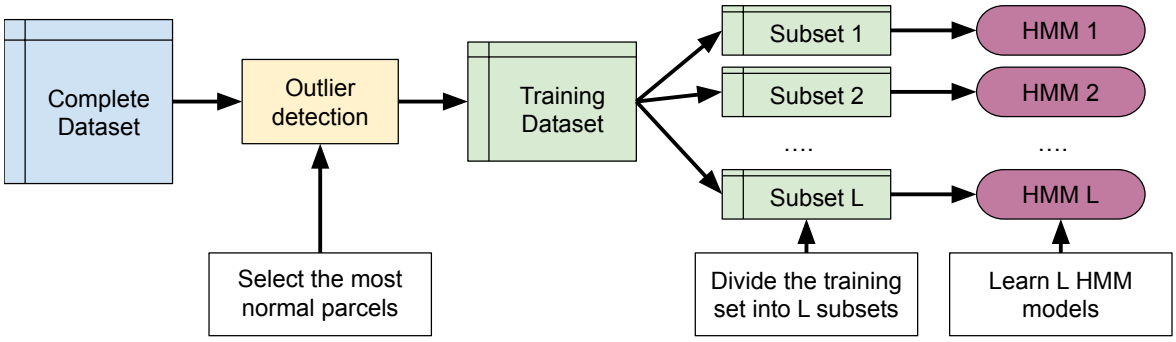


Figure 5.1: Methodological steps for learning an ensemble of HMMs.

2. **Detection at the parcel-level:** The log-probability that a parcel time series $\boldsymbol{X}^{(n)}$ has been generated by the lth HMM with parameters $\hat{\boldsymbol{\theta}}_l$ is written as

$$p_l^{(n)} = \log \mathrm{P}(\boldsymbol{X}^{(n)}|\hat{\boldsymbol{\theta}}_l), \tag{5.3}$$

which can be determined using the forward algorithm of the forward-backward procedure. The forward algorithm is provided in Algorithm 5.1, which introduces the forward variable $\alpha_t(i)$, *i.e.*, the probability that the partial observed vector $[x_1, ..., x_t]$ is in state i at time t. Using the L HMMs that have been estimated in the previous step, we

propose to attribute to each parcel $\#n$ a score denoted as $\rho^{(n)}$:

$$\rho^{(n)} = \max_{l=1,...,L} p_l^{(n)}. \tag{5.4}$$

Note that $\rho^{(n)}$ is the log-probability for the parcel $\#n$ to be associated with the most likely HMM among the L models learned during the training step. We propose to detect the abnormal parcels by comparing the score $\rho^{(n)}$ to a threshold that can then be fixed to detect a percentage of the most abnormal parcels, as with the IF algorithm. An example of the log-probabilities obtained on the rapeseed parcels is displayed in Figure 5.2. For a better visualization, these log-probabilities were scaled in the range [0,1]. Moreover, since few samples can have very large or very low probabilities, the scaling was conducted in a robust fashion using the 1th and 99th percentiles. In that example, we set a threshold to detect 10% of the most abnormal parcels.

Algorithm 5.1 Forward algorithm (pseudocode).

Input: a time series $\boldsymbol{X} = [x_1, ..., x_T]$, a HMM with D states and estimated parameters $\hat{\boldsymbol{\theta}} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$, with $a_{i,j}$ and $b_{j,t}$ the elements of the transition matrix A and the emission matrix B, respectively.

1) Initialization: $\alpha_1(i) = \pi_i b_i(x_1)$

2) Induction: $\alpha_{t+1}(i) = \left(\sum_{j=1}^{D} \alpha_t(j) a_{j,i}\right) b_i(x_{t+1})$

3) Termination: $\log P(\boldsymbol{X}|\hat{\boldsymbol{\theta}}) = \sum_{j=1}^{D} \alpha_T(j)$

Output: forward variables α, $\log P(\boldsymbol{X}|\hat{\boldsymbol{\theta}})$.

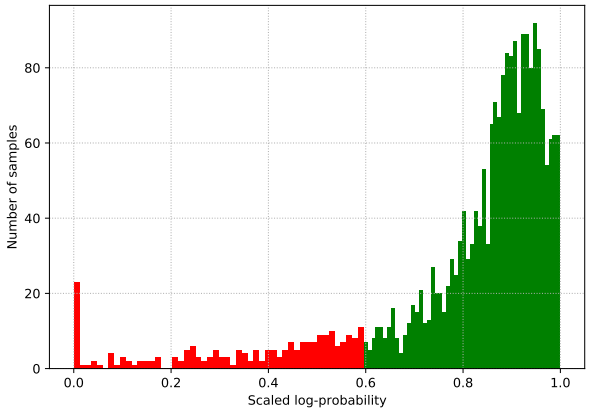


Figure 5.2: Distribution of the scaled log-probabilities attributed to each rapeseed parcel using the HMM ensemble whose hyperparemeters are provided in Table 5.1. A robust scaling of the log-probabilities was made using the 1th and 99th percentiles to have values in the range [0,1]. The separation between inliers (in green) and outliers (in red) is made by selecting 10% of the lowest probabilities.

3. **Temporal localization:** an interest of the proposed anomaly detection strategy based on HMMs is the possibility of evaluating subsequences of the analyzed time series.

Assuming that the first-order Markov chain rule stands (i.e., the current state at time t depends only on state at time $t-1$), one can evaluate the probability that the lth HMM has generated the time series $X^{(n)}$ in the temporal segment $[t_c, t_d]$ (with $c \leq d$). First, the probability of generating $X^{(n)}$ at time t is written as follows:

$$u_t = \sum_{j=1}^{D} \left\{ \sum_{i=1}^{D} \left(\alpha_{t-1}(i) a_{i,j} \right) b_j(x_t) \right\} = \sum_{j=1}^{D} \alpha_t(j). \tag{5.5}$$

Then, the log-likelihood in the time segments $[t_c, t_d]$ is defined as

$$\log \mathrm{P}(x_{t_c}, ..., x_{t_d} | \boldsymbol{\pi}, \boldsymbol{A}, b_{i,[t_c,...,t_{d-1}]}) = \log \left(\sum_{t=t_c}^{t_d} u_t \right). \tag{5.6}$$

Defining $p^{(n)}_{l,[t_c,...,t_d]} = \log \mathrm{P}(x^{(n)}_{t_c}, ..., x^{(n)}_{t_d} | \boldsymbol{\pi}^{(l)}, \boldsymbol{A}^{(l)}, b^{(l)}_{i,[t_c,...,t_{d-1}]})$ the probability that the time series of parcel $\#n$ have been generated by the lth learned HMM in the time interval $[t_c, t_d]$, we can attribute an outlier score for a given time interval as follows:

$$\rho^{(n)}_{[t_c,t_d]} = \max_{l=1,...,L} p^{(n)}_{l,[t_c,...,t_d]}, \tag{5.7}$$

which can be used to detect abnormal crop development in the interval $[t_c, t_d]$. In practice, one could choose intervals associated with a predefined growth stage, such as growing, flowering, adult-phase, senescence. For instance, to analyze the senescence stage of the rapeseed parcels, choosing the 6 last time instants of the growing season (between mid-May and early July) would be relevant. Note that if the time interval $[t_c, t_d]$ contains all the dates $t_1, ..., t_T$, we retrieve the outlier score defined in (2) for a complete growing season analysis and a single detection at the parcel-level.

5.3 Experimental results and discussion

This section validates the HMM-based anomaly detection strategy by experimental results conducted on the rapeseed parcels analyzed in Chapter 3. These results are obtained using NDVI time series (median and IQR), extracted from the 21 S2 images presented in Chapter 4 (missing features were reconstructed using the robust GMM algorithm of Chapter 4). We should highlight here that when using temporal methods, it is generally not possible to mix directly two types of data acquired at different temporal resolutions, which is the case with S1 and S2 time series. Interpolating S1 and S2 features on the same temporal grid is a possible way of solving this problem. However, we focus in this section on the median and IQR NDVI time series associated with the same S2 images, which does not require any interpolation pre-preprocessing.

The results obtained with the HMM approach are compared to those obtained with the IF algorithm (as in Chapter 3). We also consider the "Discord" algorithm (Keogh et al., 2005), which is a classical method used to detect anomalies in time series. Initially, the discord

algorithm has been designed to find the most abnormal subsequences contained in a single time series. Here, we have adapted this idea to the analysis of multiple time series. More precisely, each parcel time series is decomposed in subsequences using a sliding window of fixed length. The subsequences of all the parcel time series are then considered jointly to find the subsequences that are the farthest from their nearest neighbors. For this method, we have used the distance to the kth nearest neighbor as the outlier score attributed to each subsequence. In order to make a decision at the parcel-level, we have considered the sum among the parcel subsequences.

Finally, we would like to mention that the implementation of the HMM algorithm considered in this chapter has been made in Python, using the library *hmmlearn*[1] as a baseline. To that extent, the results presented here might slightly differ from the one presented in León-López et al. (2021), which were obtained using a MATLAB implementation based on another HMM toolbox[2].

5.3.1 Parameter tuning

- **HMM ensemble:** various hyperparameters have to be fixed to learn the ensemble of HMMs in the training step. The values of these hyperparameters chosen for the analysis are summarized in Table 5.1 and were selected by grid search. This relatively large number of parameters can be a problem in practice, when compared to other algorithms requiring simpler parameter tuning such as the IF or Discord algorithms. Nevertheless, we have observed that changing these hyperparameter values around their optimal values does not have a significant impact on the detection results.

Table 5.1: Hyperparameters used to learn the ensemble of HMMs.

Hyperparameter	Value
Number of states in each HMM	18
Number of Gaussians for the emission distribution	13
Number of models L	10
Size of the training dataset N_{training}	500
Subsampling N_s	100

- **Discord:** the number of k-nearest neighbors and the length of the sliding windows are the only hyperparameters to choose when using the Discord algorithm. A value of $k = 100$ was chosen (varying k in the range $[10, 1000]$ did not have a significant impact on the detection results). The length of the sliding window was fixed to 3 by cross validation (choosing close values leads to similar results).

- **IF algorithm:** the IF algorithm was used with the hyperparameters fixed in Chapter 3 (the number of trees is $n_{\text{trees}} = 1000$ and the subsampling parameter is $n_{\text{samples}} = 256$).

[1] https://github.com/hmmlearn/hmmlearn
[2] https://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

5.3.2 Detection results on the rapeseed parcels

Using the rapeseed dataset presented in Chapter 2, outlier scores were attributed to the parcels by the 3 different algorithms (HMM, Discord and IF). Precision versus outlier curves are displayed in Figure 5.3. It can be observed that similar area under the precision-recall curve (AUC) can be reached with the IF and HMM approaches: AUC = 0.89 for the IF algorithm, while it is slightly lower when using the HMM approach (AUC = 0.86). The difference in AUC is explained by a lower precision obtained with the HMM algorithm for outlier ratios higher than 10%, which means that the strongest anomalies are detected with both algorithms. On the other hand, the Discord algorithm provides a lower AUC (equal to 0.80). The small values of AUC obtained when using the Discord approach can mainly be explained by two factors. First, using sliding windows leads to compare subsequences delayed in time, which prevents some anomalies such as delayed senescence to be detected. Second, the Discord algorithm uses the Euclidean distance as an outlier score, which seems to be not appropriate for this application. More relevant metrics such as those used in the IF algorithm might be investigated to improve the detection results. Finally, a further investigation showed that when choosing an outlier ratio equal 10%, all the three approaches detect similar proportions of parcels within each outlier category, as depicted in Figure 5.4.

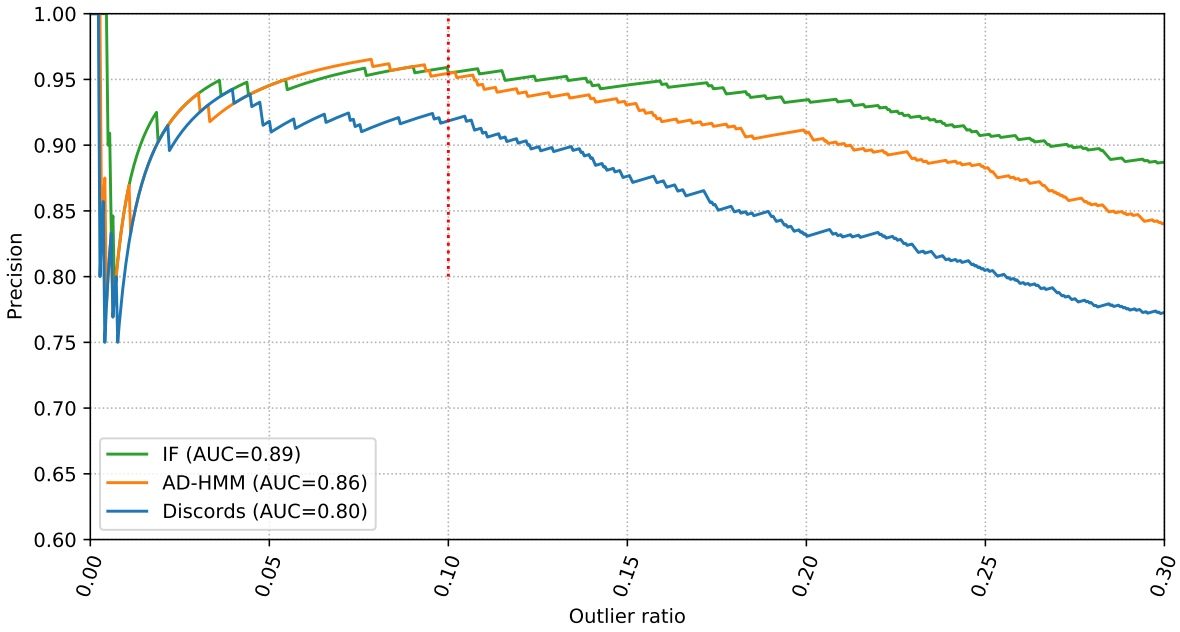


Figure 5.3: Precision vs. outlier ratio curves for the rapeseed dataset using the IF (green), the ensemble of HMMs (orange) and the Discords (blue) algorithms.

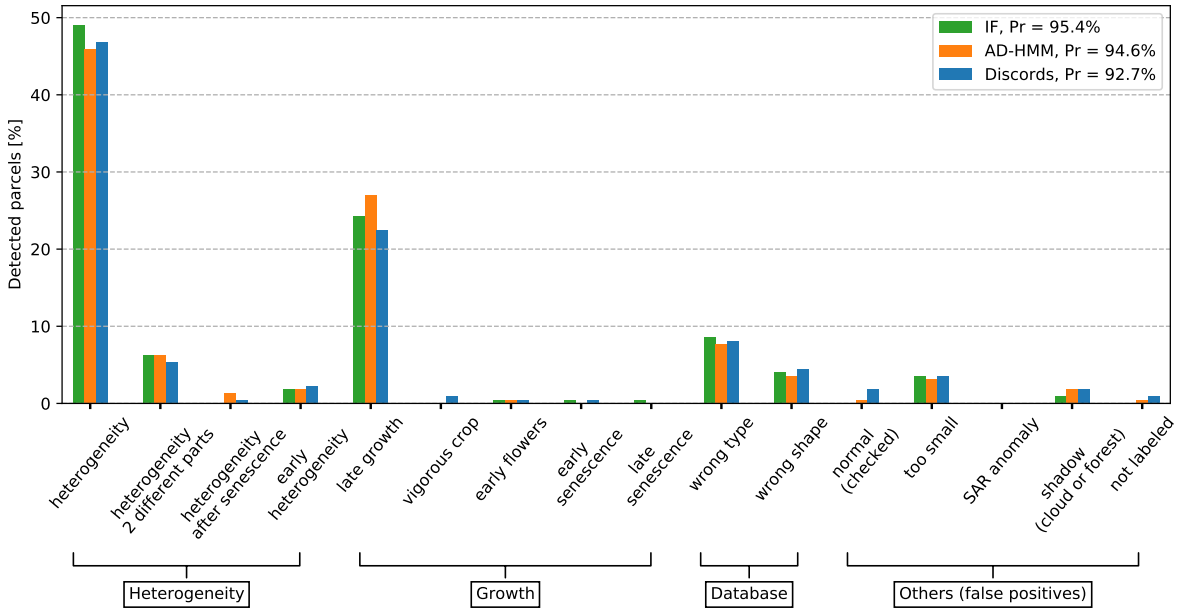


Figure 5.4: Distribution of the detected rapeseed parcels within the different outlier categories using the IF (green), the ensemble of HMMs (orange) and the Discords (blue) algorithms. The detection is made with an outlier ratio equal to 10%.

5.3.3 Anomaly localization

Figure 5.5(a, c,e) show three examples of anomaly localization obtained using the score of Equation 3. For clarity, these figures show the most informative indicator for each analyzed parcel (i.e., median NDVI or IQR NDVI). They were obtained using an outlier ratio equal to 10% and 5 different temporal intervals covering the growing season (delimited by the vertical gray lines). One can appreciate that the most abnormal parts of the growing season are detected as anomalous (in red). Figure 5.5(a) was labeled as "late growth", Figure 5.5(c) as "early flowering/senescence" and Figure 5.5(e) as "heterogeneity after senescence".

Figure 5.5(b,d,f) provide the forward log-probabilities attributed by each of the 10 HMMs to the parcels analyzed in Figure 5.5(a,c,e). These probabilities are computed using Equation 3. One can see the interest of using an ensemble of HMMs to capture the different possible behaviors of the crop parcels. For instance, one can see that the behavior of parcel (a) is better captured by models 4 and 7. Moreover, looking at the induction step of Algorithm 5.1, one can notice that the forward probabilities computed at time $t+1$ depends on the forward probabilities computed at time t. In practice, this means that for "normal" subsequences, the forward log-probabilities tend to increase through time (*i.e.*, each new "normal" observation increases the probability that the time series has been generated by the HMM). On the contrary, the log-probabilities tend to decrease when new observations do not conform to the learned HMMs. This behavior can be observed for the parcel analyzed in Figure 5.5(c), where its log-probabilities (Figure 5.5(d)) increase during the first part of the growing season, and tend to decrease after winter.

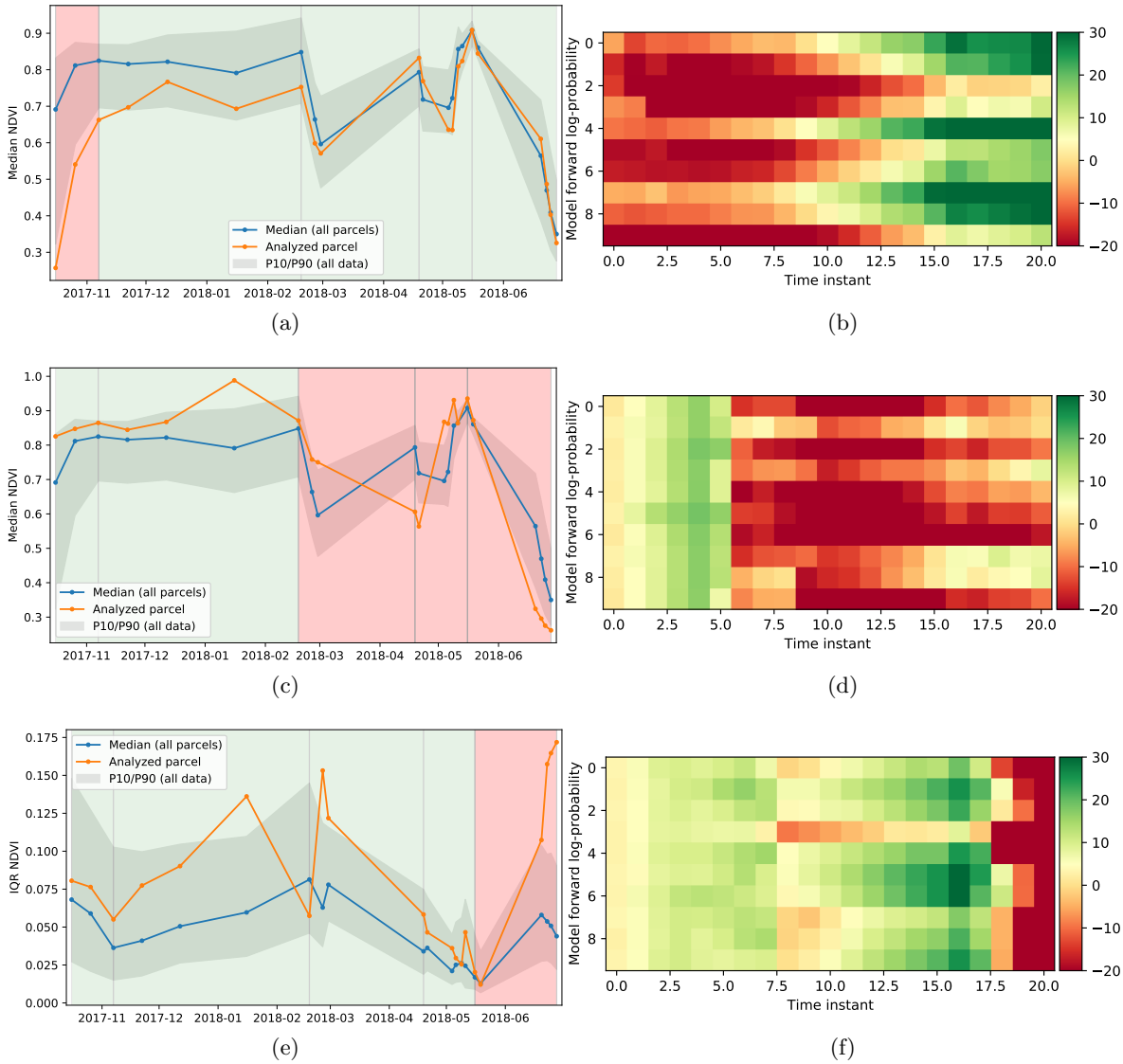


Figure 5.5: Examples of time series for 3 rapeseed parcels are displayed in (a,c,e). In each figure, the blue line represents the median value of the whole dataset, whereas the orange line corresponds to the time series of the analyzed parcel. The shaded area is filled between the 10th and 90th percentiles. Areas in green were not detected as anomalies whereas areas in red were detected as anomalies (the vertical gray lines delimitate the different temporal segments considered). Forward log-probabilities attributed by each of the 10 HMM to the parcels analyzed in figure (a), (c) and (e) are displayed in figures (b), (d) and (f), respectively.

5.4 Conclusion and perspectives

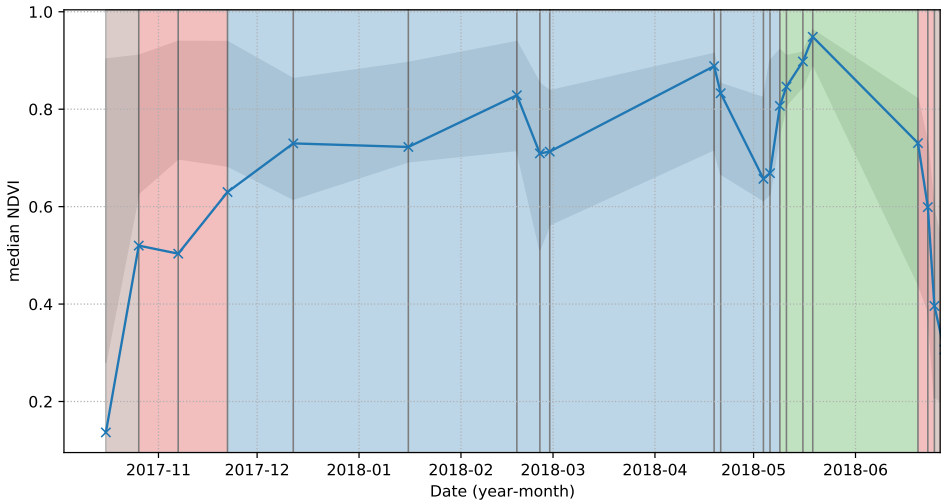
This chapter is a preliminary work exploring temporal approaches for the detection and temporal localization of anomalies in agricultural parcels. We proposed a method for detecting anomalies in the development of crop parcels based on an ensemble of HMMS. The main idea of this method is to model the underlying behavior of the crop parcels with various HMMs, and detect the crop parcels that have phenological behaviors differing significantly from these models. An advantage of the proposed approach is that it can be conducted on particular temporal segments of the growing season, to localize the occurrence of the anomalies. These temporal segments can be chosen regularly throughout the growing season or can be fixed by the user (e.g., to match the phenological stages of the growing season).

The HMM-based approach is not significantly improving the results (in terms of precision/recall) when compared to simpler methods such as the IF algorithm. Moreover, the localization of anomalies is also possible with IF using for instance the SHAP method, which exploits the anomaly scores provided by the algorithm. However, once a HMM has been learned, it can be used on new sequences of various lengths. An interesting application could be for instance to learn HMMs on a given growing season, and then use the models on a new growing season to predict potential anomalies. This could be particularly useful if few samples are available for the new growing season (note that the IF algorithm cannot be used directly for this usecase).

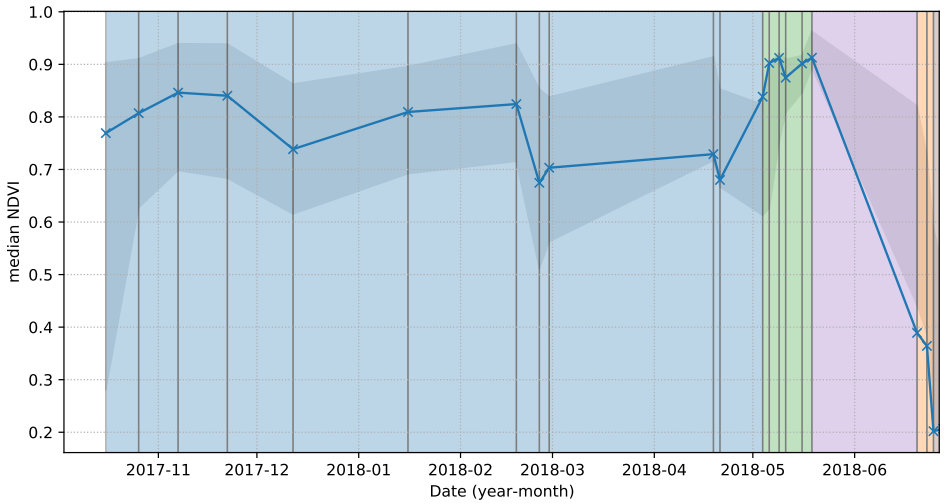
Various other perspectives would be interesting to study in future work. First, adapting the HMM approach to allow the use of multiple sensors (*e.g.*, S1 and S2 data) would be relevant, particularly when looking at the results presented in Chapter 3, which showed that using additional S1 data improves performance for detecting anomalies in rapeseed crops. Promising results were obtained after interpolating S1 and S2 time series on the same temporal grid. However, more tests should be conducted, e.g., to determine the best interpolation method for this application.

Finally, we think that other models dedicated to the clustering of time series in different phenological stages would deserve to be investigated. The idea is to analyze directly subsequences of time series to capture more efficiently the intrinsic relationships between the different time instants of the growing season. This problem can be addressed by using clustering methods such as the Toeplitz Inverse Covariance-Based Clustering (TICC), which was originally proposed in Hallac et al. (2017). In brief, the TICC method aims at clustering subsequences of multivariate time series within a GMM framework. An originality of this approach is to perform the clustering based on a graphical dependency structure for each subsequence (i.e., by imposing a Toeplitz inverse structure to the covariance matrices). Moreover, a temporal consistency constraint is added to encourage adjacent subsequence to be part of a same cluster. Promising results were already obtained by modifying the TICC method to adapt it to crop monitoring, in particular to handle multiple time series coming from different parcels. In addition to the detection and localization of anomalies, an interesting byproduct of this approach is to automatically find for each parcel the different stages of the growing season (e.g., growth, flowering, senescence, etc.). An example of the clustering obtained for

two different rapeseed parcels is presented in Figure 5.6 (the different colors correspond to different the stages attributed by the algorithm). In Figure 5.6(a), the late growth of the parcel is captured and two new states are attributed to the early season when compared to the state of the parcel presented in Figure 5.6(b). Moreover, one can see that the delay in the flowering stages of the two parcels is also captured. Further investigations should be done to validate these first results and their relevance for crop monitoring.



(a)



(b)

Figure 5.6: Clustering obtained for two specific rapeseed parcels using our modified TICC algorithm (the model is learned on the whole dataset with a number of clusters set to 6). The parcel in (a) is affected by a late growth, and the parcel (b) has an early senescence. The blue line correspond to the median of the NDVI time series of the analyzed parcel, while the shaded area is filled between the 10th and 90th percentiles of the whole dataset. The different colored rectangles correspond to different states attributed by TICC.

Chapter 6

Conclusions

The objective of this thesis was to study new crop monitoring methods at the parcel-level using data extracted from multispectral and SAR satellites (such as Sentinel 1 and Sentinel 2 satellites). A particular attention has been given to the automatic detection of anomalous crop development, a problem that has received few attention in the literature.

In Chapter 1, we have introduced the interest of remote sensing for agriculture. A focus has been made on S1 and S2 satellites, which are particularly well suited for crop monitoring at the parcel-level thanks to their spatial and temporal resolutions. Using data coming from two types of satellites (synthetic aperture radar for S1 and multispectral imagery for S2) was motivated by their complementary for the analysis of crop parcels. This chapter has also introduced the problem of anomaly detection in the vegetation status and its interest for the monitoring and optimization of agricultural practices. A state-of-the-art has motivated the need for new strategies adapted to the detection of anomalous crop development at the parcel-level, in particular when analyzing a single growing season (or a part of a growing season).

In Chapter 2, we have detailed the different processing steps for the extraction of parcel-level features coming from S1 and S2 data. This feature extraction is decomposed into 3 main steps: 1) preprocessing of S1 and S2 images, 2) computation of pixel-level features and 3) computation of spatial statistics at the parcel-level from the pixel-level features associated with different dates. The preprocessing of remote sensing data is a classical step needed before feature extraction. For our use case, one can mention that a terrain flattening operation was added to the SAR processing chain to take into account the differences in soil geometry. The pixel-level features recommended for crop monitoring are vegetation indices for S2 data (in particular the NDVI) and backscattering coefficients VV and VH for S1 data. Throughout this manuscript, we have shown their relevance and their complementary for crop monitoring, as well as their ease of interpretation. To extract the parcel-level features, we have proposed to use spatial statistics computed using the parcel boundaries. It has been shown that the median and interquartile are two relevant statistics that can summarize efficiently and robustly the pixel-level features in terms of mean value and heterogeneity.

In Chapter 3, we have proposed a fully unsupervised strategy to detect anomalous crop development at the parcel-level using the features computed in Chapter 2. The core of this strategy is based on outlier detection algorithms, which are designed to find samples that are not conform to the majority of samples contained in the dataset. Overall, we showed that the isolation forest algorithm is a good choice for crop monitoring since it is fast, robust to changes

in the dataset or in the crop type and does not require an extensive hyperparameter tuning. This chapter allowed us to meet various important objectives of the thesis. In particular, we have shown that it is possible to detect anomalous crop development in various crop types by analyzing a single (or a part of a) growing season. Moreover, we have shown the interest of using SAR and multispectral data jointly to improve the detection results. Finally, the outlier score provided by the isolation forest algorithm is generally higher for strong anomalies (*e.g.*, errors related to the crop type), which is interesting to sort the parcels by their degree of abnormality.

Chapter 4 addressed an important challenge related to the presence of missing data in the parcel time series. We have proposed to impute missing data using a Gaussian Mixture Model (GMM), whose parameters are estimated by the Expectation-Maximization (EM) algorithm. All other tested methods were outperformed by this approach, which achieves competitive reconstruction errors even with a small dataset (containing around two thousand samples). In addition, we have proposed a novel strategy for the robust estimation of GMM parameters, which is useful when the dataset is contaminated by strong outliers (*e.g.*, parcels coming from a different crop type than the one analyzed). The proposed robust GMM makes use of the isolation forest algorithm to attribute weights to the different parcels of the database, which leads to a data imputation and an anomaly detection within the same EM algorithm. We have shown that the detection of anomalous crop development in the presence of missing data can be conducted with good accuracy with this strategy. Finally, it has been shown that using additional S1 data can help improve both reconstruction and detection results.

In Chapter 5, we have explored new anomaly detection strategies taking into account temporal correlations of the times series associated with crop parcels. More precisely, we have focused on an ensemble of hidden Markov models (HMMs) adapted to detect and localize anomalies in multivariate time series. We have shown that this approach can lead to detection results similar to those obtained with the point outlier detection algorithms tested in Chapter 3, with the advantage of allowing anomaly localization. For general applications such as the one presented in this thesis, using the method proposed in Chapter 3 is recommended since it is faster and less dependent on the choice of the hyperparameters. However, the algorithm investigated in this chapter is interesting for applications requiring anomaly localization. Finally, we have presented preliminary results on time series clustering with GMMs. This approach could be interesting to automatically detect the different phenological stages of a crop parcel.

The results presented in this thesis opens several interesting perspectives that should be investigated in future work.

- **Toward an operational service:** a direct and important perspective is to adapt the proposed method to an operational context. The detection of anomalous parcels could for instance be an additional information provided by a crop monitoring service. The main challenges for this implementation are related to automatic data processing and feature extraction. Other challenges are related to the user experience, especially to have a user-friendly interpretation of the results. In that context, an interesting long-term perspective could be to incorporate user feedbacks to the anomaly detection results. This could be for instance done with the isolation forest algorithm, as explained in Das et al. (2016). The main difficulty of this task is to evaluate the performance and the quality of this user feedback. Another difficulty is to find representative examples that can minimize user interactions. Finally, the data imputation method proposed in this thesis could be used in an operational context in a wider range of applications related to vegetation monitoring (*e.g.*, for the monitoring of vine production[1]). For such applications, the number of parcels available for analysis could be a main challenge.

- **Classification of the outlier parcels:** a perspective directly related to the previous point is to be able to classify the anomalies affecting the detected parcels. A short-term perspective is to provide tools to help the user to identify the different types of anomalies. A long-term (and much more difficult) perspective is the automatic classification of the outlier parcels into different categories. First attempts for a supervised classification of anomalies have been made in León-López et al. (2021). However, they would deserve to be further studied for operational services. The main challenges related to this task are 1) the relatively small number of anomalies annotated by experts, 2) the variety of the anomaly types and 3) the fact that outlier parcels can belong to several categories simultaneously (i.e., an anomalous parcel can be affected by late growth and heterogeneity). Finally, it would not be realistic to label each new dataset from an operational point of view.

- **Crop types:** an obvious future work is to apply the proposed method to other crop or vegetation types. First attempts have been made for vineyards, with promising results (these results have not been reported in this PhD thesis for brevity). We have observed good reconstruction results with robust GMMs, even with small size datasets (Mean absolute reconstruction error of the median NDVI around 0.018 in a databased composed of 1200 parcels). For vineyards, using a GMM is particularly interesting, since this type of vegetation can have a wide range of different phenological developments within a same growing season (*e.g.*, the inter-row surface can be covered by grass). A main challenge related to this crop type is the relatively small size of the parcels, which can be problematic with the spatial resolution of Sentinel data.

- **Features:** other pixel-level features could be investigated and compared to the ones presented in this manuscript. Obvious examples are biophysical indicators extracted

[1] http://oenoview365.terranis.fr/

from multispectral images (*e.g.*, leaf area index) or indicators extracted from single look complex (SLC) SAR images (*e.g.*, phase coherence, entropy). Depending on the final application, choosing the features and the spatial statistics can be a challenge, since they are directly correlated to the type of anomalies detected.

- **Pixel-level analysis:** another perspective could be to propose crop monitoring methods applicable at the pixel-level. We should first highlight that working at the parcel-level has multiple advantages, in particular 1) it reduces computing and storage requirements, which is a common issue in remote sensing applications (Inglada et al., 2017) and 2) it allows us to work with SAR images without further processing to reduce the speckle noise, which is also computationally intensive. On the other hand, working at the pixel-level would increase the number of samples to be analyzed, which could be interesting to fight against the curse of dimensionality. Having a decision at the pixel-level could also be interesting to localize spatially the areas with anomalous development.

- **Various growing seasons:** an interesting long-term perspective is to extend the problem studied in this manuscript to the analysis of multiple growing seasons. Various challenges should be addressed in that situation, as for instance the problem of crop rotation (i.e., the crop type associated with a parcel changes with time), the problem of temporal inconsistencies or the challenge caused by inter-annual variability (using growing degree days could be for instance important for such application). In that context, more investigations should be conducted to analyze the potential interest of temporal approaches, as for instance to analyze a small number of parcels coming from a new growing season.

- **GMM Regularization:** we have found that imputing missing data with GMMs is well suited to crop monitoring. However, we think that these results could be further improved by constraining the structure of the covariance matrices, especially when working with high dimensional data. We should remind here that various attempts have been made during this PhD thesis without success (*e.g.*, adding a sparsity constraint on the precision or covariance matrices). One interesting possibility is to impose a Toeplitz constraint on the covariance matrices. While we have obtained promising results on synthetic datasets, it is not the case for real-world experiments, mainly because of the reduced temporal resolution of S2 data, which breaks the Toeplitz structure of the covariance matrices. Using SAR data could be a way to address this issue. Moreover, all our experimentations were made using a unique Toeplitz structure. Using multiple features requires more theoretical work, e.g., allowing block-Toeplitz covariance matrices to be considered.

Chapter 7

Résumé de la thèse en français

Contents

7.1 Introduction

7.1.1 Contexte général

L'agriculture devra nourrir plus de 10 milliards de personnes d'ici 2050, augmentant la demande agricole de 50% par rapport à 2013 dans un scénario de croissance modeste (FAO, 2017). Cet essor de la production alimentaire va se confronter au changement climatique, qui impactera la sécurité alimentaire à différents niveaux de la chaîne de production alimentaire (Tirado et al., 2010; Wheeler and von Braun, 2013). En outre, un changement des pratiques agricoles est nécessaire pour diminuer leurs impacts négatifs sur la biodiversité (Newbold et al., 2015), les ressources en eau et les émissions de gaz à effet de serre (Gomiero et al., 2011).

Dans ce contexte, la surveillance de la croissance et de l'état des cultures agricoles devient un enjeu nécessaire s'adressant à un grand nombre d'acteurs (Weiss et al., 2020). La télédétection peut fournir des informations essentielles au secteur agricole en temps opportun et de manière fiable, et ce, à grande échelle (Atzberger, 2013). Elle est donc importante pour mesurer l'intensification durable de la production agricole et optimiser les pratiques culturales (Areal et al., 2018). De plus, la surveillance en temps quasi-réel peut contribuer à améliorer la résilience du système alimentaire et à réagir aux événements extrêmes (Wheeler and von Braun, 2013).

Les différentes applications de la télédétection pour l'agriculture peuvent être regroupées en 4 catégories : phénotypage, prévision du rendement, services écosystémiques et agriculture de précision (Weiss et al., 2020). Cette dernière catégorie, qui fait l'objet de cette thèse, vise à surveiller les cultures pour optimiser les rendements ainsi que les pratiques agricoles. Elle couvre un large éventail d'applications, comme la détection de mauvaises herbes et de maladies (López-Granados, 2011; Mahlein, 2016) et la surveillance des nutriments et du stress hydrique (Baret et al., 2007; Calera et al., 2017). L'utilisation des images issues de la télédétection est particulièrement intéressante pour l'agriculture de précision car elles fournissent des informations spatiales et temporelles sur l'état des cultures, et ce, de manière non destructive et sans nécessiter de visites sur place (Schulz et al., 2021).

7.1.2 Utilisation des satellites Sentinel-1 et Sentinel-2 pour l'agriculture de précision

Historiquement, les applications de la télédétection pour l'agriculture étaient principalement limitées par les capteurs utilisés dans les satellites (en particulier, leur gamme spectrale et leur résolution spatiale) et par le temps de revisite de ces derniers (Moran et al., 1997). Depuis quelques années, la quantité d'images de télédétection librement accessibles a considérablement augmenté, en particulier grâce à la mission Copernicus de l'Union européenne (UE), opérée par l'Agence spatiale européenne (ASE). Son premier satellite multispectral à haute résolution (Sentinel-2A) a été lancé en 2015, suivi par un deuxième en 2017 (Sentinel-2B)

(Drusch et al., 2012). Deux satellites radar à synthèse d'ouverture (RSO ou SAR en anglais), Sentinel-1A et Sentinel-1B, ont été lancés respectivement en 2014 et 2016 (Torres et al., 2012). Les satellites Sentinel-1 (S1) et Sentinel-2 (S2) ont une haute résolution temporelle et spatiale qui sont adaptées pour travailler au niveau de la parcelle (pour une analyse à très haute résolution, par exemple au niveau de la plante, les résolutions ne sont cependant pas suffisantes). Les deux types de capteurs sont complémentaires et ont été largement étudiés dans ce contexte. Tous ces facteurs (libre accès, résolutions temporelle et spatiale, caractéristiques adaptées au suivi des cultures agricoles) ont motivé l'utilisation des satellites S1 et S2 dans le cadre de cette thèse.

7.1.3 Détection d'anomalies dans la végétation : état de l'art

Un enjeu peu étudié en agriculture de précision est la détection automatique de parcelles agricoles présentant un développement phénologique anormal. A titre d'exemple, la Figure 7.1 permet de visualiser plusieurs champs agricoles à l'aide d'images S2 acquises (a) le 25 février 2018 et (b) le 21 avril 2018. A ce stade de la saison de croissance (fin de l'hiver / début de la floraison), on peut remarquer que certaines parcelles agricoles (ici de colza) sont plus ou moins affectées par des hétérogénéités, et que celles-ci peuvent être plus ou moins transitoires. La détection de parcelles dont le comportement phénologique diffèrent significativement des autres pourrait aider des parties prenantes tels que les agriculteurs ou les coopératives agricoles à optimiser les pratiques agricoles, détecter des maladies ou optimiser la fertilisation des parcelles. Elle pourrait également être utile dans des domaines tels que le contrôle des subventions (notamment dans le cadre de la politique agricole commune en Europe) ou l'assurance-récolte.

(a) (b)

Figure 7.1: Parcelles de colza (contours rouges et jaunes) visualisées avec des images S2 acquises (a) le 25 février 2018 et (b) le 21 avril 2018.

Dans le domaine de l'observation de la Terre, la plupart des études se sont concentrées sur la détection d'anomalie dans la végétation dans le cadre d'analyses à grande échelle (au niveau

d'une région ou d'un pays). Dans la grande majorité des cas, ces études utilisent des séries temporelles construites à partir de l'indice différentiel normalisé de végétation (*Normalized Difference Vegetation Index* ou NDVI en anglais). La plupart des approches proposées sont dites *prédictives* (Chandola et al., 2009; Aggarwal, 2017). Ces techniques cherchent d'abord à modéliser les séries temporelles de NDVI, dans un second temps elles utilisent ce modèle pour détecter de potentielles anomalies en comparant les nouvelles observations avec les valeurs prédites.

Les approches les plus courantes sont basées sur des modèles paramétriques décrivant l'évolution temporelle du NDVI (Atzberger and Eilers, 2011a; Beck et al., 2006) ou des techniques de filtrage (ou lissage) (Atzberger and Eilers, 2011b; Hird and McDermid, 2009; Klisch and Atzberger, 2016; Meroni et al., 2019). De nombreuses autres approches ont été étudiées, comme par exemples les processus autorégressif *SARIMA* (*Seasonal Autoregressive Integrated Moving Average*) (Zhou et al., 2016) ou les filtres de Kalman (*extended Kalman filter)* (Sedano et al., 2015). Plus récemment, des techniques similaires (*Breaks for Additive Season and Trend*, BFAST) ont été utilisées avec des données S2, par exemples pour détecter des anomalies liées à l'utilisation des sols dans le cadre de la PAC (Kanjir et al., 2018). Notons que la technique BFAST a été d'abord introduite par Verbesselt et al. (2010) pour contrôler la détection des changements phénologiques dans les séries temporelles de NDVI.

Les approches mentionnées précédemment peuvent être difficiles à mettre en œuvre pour notre cas d'utilisation, qui consiste à détecter un développement anormal dans les parcelles d'un type de culture donné. Premièrement, la modélisation du comportement normal des données implique d'avoir accès à des exemples représentatifs normaux, ce qui peut s'avérer difficile et laborieux en pratique. La rotation des cultures, le manque de données de référence et la parcimonie des séries temporelles S2 causée par la couverture nuageuse sont d'autres facteurs qui rendent cette mise en œuvre encore plus difficile. Ceci est d'autant plus problématique car les techniques prédictives ont généralement besoin de longues séries temporelles de référence pour être calibrées efficacement. Dans notre cas, l'analyse d'une seule saison de croissance est plus pertinente, principalement pour des raisons de coût et de mise en oeuvre opérationnelle. Avoir accès à des données parcellaires fiables provenant de plusieurs saisons de croissance pour construire des modèles temporels serait également problématique en pratique. Enfin, tandis que la plupart des études se concentrent uniquement sur l'analyse du NDVI, il semble pertinent d'utiliser une plus grande variété d'indicateurs provenant de données S1 et S2 afin de mieux caractériser les champs agricoles. Ce bref état de l'art montre qu'il est nécessaire d'étudier de nouvelles approches de détection d'anomalies dédiées spécifiquement à la surveillance des parcelles agricoles.

7.1.4 Formulation du problèmes et objectifs de la thèse

Cette thèse à pour but d'explorer les défis liés à la surveillance automatique des parcelles agricoles à partir de données S1 et S2. En particulier, un enjeu principal est de détecter les parcelles agricoles dont le comportement phénologique diffère significativement des autres. Une hypothèse est que les contours de la parcelle et le type de culture sont disponibles,

l'analyse étant conduite pour un type de culture agricole donné. L'analyse temporelle a été limitée à une seule saison de croissance, principalement pour des raisons de faisabilités (comme expliqué dans la section précédente). Pour les mêmes raisons, la méthode proposée doit être la moins supervisée possible (par exemple, en ce qui concerne le réglage des paramètres de l'algorithme de détection ou la nécessité d'avoir à disposition des données étiquetées). Un autre défi important, qui est récurrent en télédétection, est la prise en compte des données manquantes (provenant des nuages pour les images S2 ou des problèmes d'acquisition). Les principaux objectifs et défis de cette thèse sont résumés dans ce qui suit :

→ Détecter des anomalies pertinentes, c'est-à-dire liées à un phénomène agronomique, au niveau de la parcelle.

→ Attribuer un score d'anomalie proportionnel au degré d'anormalité de la parcelle.

→ Être capable d'analyser les parcelles au cours d'une seule saison de croissance (ou d'une partie de la saison de croissance si possible)

→ Utiliser efficacement la complémentarité des données S1 et S2.

→ Valider la méthode sur différents types de cultures.

→ Proposer une méthode entièrement automatisée (sans besoin d'étiquetage manuel et de réglage des paramètres).

→ Traiter les données manquantes causées par des nuages ou des problèmes d'acquisition.

7.2 Résumé du Chapitre 2 : pré-traitement des données pour l'extraction d'indicateurs au niveau parcelle

Ce chapitre présente la zone d'étude, les données parcellaires et les données de télédétection utilisées dans les différentes expériences menées tout au long de la thèse. Une attention particulière est consacrée au traitement des données issus des satellites S1 et S2. Ce pré-traitement est justifié principalement par deux raisons : 1) le pré-traitement des données peut améliorer la qualité des images de télédétection (par exemple, correction des effets atmosphériques pour les images S2 (Hagolle et al., 2015), calibration et correction du terrain pour les images S1), 2) disposer d'éléments dont l'interprétation est facilitée peut améliorer les résultats et, surtout, faciliter leur interprétation. De plus, l'extraction d'indicateurs le plus pertinents possible est généralement recommandée dans le cadre d'analyses non supervisées.

La chaîne de traitement proposées se décompose en 3 étapes principales : 1) prétraitement des images S1 et S2, 2) extraction d'indicateurs au niveau du pixel pour chaque image et 3) calcul de statistiques spatiales au niveau de la parcelle pour chaque indicateur en (2). Le prétraitement des données de télédétection est une étape classique effectuée avant l'extraction de caractéristiques au niveau pixel. Notons que dans notre cas, une opération d'aplatissement du terrain (*terrain flattening* en anglais) a été ajoutée à la chaîne de traitement des données S1 pour prendre en compte les différences de géométrie du sol. Les caractéristiques au niveau pixel recommandées pour le suivi des cultures sont les indices de végétation pour les données S2 (en particulier le NDVI) et les coefficients de rétrodiffusion VV et VH pour les données S1. Tout au long de ce manuscrit, nous avons montré leur pertinence et leur complémentarité pour la surveillance des parcelles agricoles, ainsi que leur facilité d'interprétation. Pour extraire les caractéristiques au niveau des parcelles, nous avons proposé d'utiliser des statistiques spatiales calculés à partir des contours des parcelles. Nous avons montré que la médiane et l'écart interquartile sont deux statistiques pertinentes qui peuvent résumer de manière efficace et robuste le comportement d'un indicateur agronomique en termes de valeur moyenne et d'hétérogénéité.

7.3 Résumé du Chapitre 3 : détection de parcelles agricoles anormales à l'aide de séries temporelles multispectrales et RSO : application au blé et au colza

Dans ce chapitre, nous avons proposé une stratégie non supervisée dédiée à la détection de parcelles agricoles qui ont un développement phénologique anormal. Cette méthode utilise en entrée les indicateurs calculées dans chapitre précédent. La stratégie proposée est basée sur des algorithmes de détection d'anomalie, spécialement conçus pour trouver des observations différant de manière significative de la majorité des autres données. Dans l'ensemble, nous avons montré que l'algorithme de forêt d'isolement (*Isolation Forest* en anglais) est le mieux adapté à notre cas d'usage. En effet, il est rapide, peu sensible aux changements (par exemple concernant les indicateurs utilisés ou le type de culture analysée) et ne nécessite pas un réglage

approfondi de ses hyperparamètres (ce point est particulièrement important dans un contexte non-supervisé). Ce chapitre nous a permis d'atteindre plusieurs objectifs importants de la thèse. En particulier, nous avons montré qu'il est possible de détecter un développement anormal dans différents types de cultures, et ce, en analysant une seule (ou une partie d'une) saison de croissance. De plus, nous avons montré l'intérêt d'utiliser conjointement des données RSO et multispectrales pour améliorer les résultats de détection. Enfin, le score d'anomalie fourni par l'algorithme de forêt d'isolement est en moyenne plus élevé pour les anomalies sévères (par exemple, causées par une erreurs liées au type de culture reporté dans la base de donnée), ce qui est intéressant pour trier les parcelles selon leur degré d'anormalité.

7.4 Résumé du Chapitre 4 : reconstruction de séries temporelles Sentinel-2 avec données manquantes à l'aide de modèles de mélange gaussien

Ce chapitre a abordé un défi important lié à la présence de données manquantes dans les séries temporelles associées aux parcelles agricoles. Nous avons proposé de reconstruire les données manquantes en utilisant des modèles de mélange gaussien (*Gaussian Mixture Model (GMM)* en anglais), dont les paramètres sont estimés par l'algorithme espérance-maximisation (EM). Toutes les autres méthodes testées ont été significativement moins performantes que cette approche, qui permet d'obtenir de faibles erreurs de reconstruction, et ce, même avec une base de données de taille limitée (contenant tout au plus deux mille parcelles). De plus, nous avons proposé une nouvelle stratégie pour l'estimation robuste des paramètres du GMM. Ceci est utile lorsque l'ensemble de données est contaminé par de fortes valeurs anormales (par exemple, à cause de parcelles provenant d'un type de culture différent de celui analysé). L'estimation robuste du GMM qui est proposée utilise l'algorithme de forêt d'isolement pour attribuer des poids aux différentes parcelles de la base de données (les parcelles anormales ont un poids réduit, voire nul). Ceci conduit à une imputation des données manquantes et à une détection d'anomalies au sein du même algorithme EM. Nous avons montré que la détection de parcelles au développement anormal en présence de données manquantes peut être réalisée avec une bonne précision grâce à cette stratégie. Enfin, nous avons également mis en avant l'intérêt d'utiliser des données S1 supplémentaires pour améliorer les résultats de reconstruction et détection.

7.5 Résumé du Chapitre 5 : vers des approches temporelles pour la détection et la localisation de développement anormal dans les parcelles agricoles

Dans ce chapitre, nous avons exploré l'intérêt d'utiliser des approches temporelles pour la détection et la localisation de développement anormal dans les parcelles agricoles. Plus précisément, nous nous sommes concentrés sur l'utilisation d'un ensemble de modèles de Markov

cachés (MMC ou*Hidden Markov model (HMM)* en anglais) adapté pour détecter et localiser des anomalies dans des séries temporelles multivariées. Nous avons montré que cette approche peut conduire à des résultats de détection similaires à ceux obtenus avec les algorithmes de détection testés dans le Chapitre 2, avec l'avantage de permettre la localisation des anomalies. Pour des applications générales telles que celle présentée dans cette thèse, l'utilisation de la méthode proposée dans le Chapitre 2 est toutefois recommandée car elle est plus rapide et moins sensible au choix des hyperparamètres. Cependant, l'algorithme étudié dans ce chapitre est intéressant pour les applications nécessitant la localisation d'anomalies. Enfin, nous avons présenté des résultats préliminaires sur le partitionnement (ou *clustering* en anglais) de séries temporelles avec des GMMs. Cette approche pourrait par exemple être intéressante pour détecter automatiquement les différents stades phénologiques d'une parcelle de culture.

7.6 Conclusion et perspectives

L'objectif de cette thèse était d'étudier de nouvelles méthodes de suivi des cultures au niveau de la parcelle en utilisant des données extraites de satellites multispectraux et RSO (tels que les satellites Sentinel 1 et Sentinel 2). Une attention particulière a été accordée à la détection automatique de développement anormal dans les cultures, un problème qui a reçu jusqu'à présent peu d'attention dans la littérature.

Les résultats obtenus tout au long de cette thèse ouvrent plusieurs perspectives intéressantes qui devraient être étudiées dans des travaux futurs.

- **Vers un service opérationnel :** une perspective directe et importante est d'adapter la méthode proposée à un contexte opérationnel. La détection de parcelles anormales pourrait par exemple être une information supplémentaire fournie par un service de surveillance des cultures. Les principaux défis pour cette mise en œuvre sont liés au traitement automatique des données. D'autres défis concernant l'expérience de l'utilisateur devront être relevés, notamment pour faciliter l'interprétation des résultats. Une perspective intéressante à long terme pourrait être de prendre en compte les retours des utilisateurs lors de la détection des anomalies. Cela pourrait être fait par exemple avec l'algorithme de forêt d'isolement, comme proposé dans Das et al. (2016). La principale difficulté de cette tâche est d'évaluer la performance et la qualité de la prise en compte du retour utilisateur. Une autre difficulté est de trouver des exemples représentatifs afin de minimiser les interactions de l'utilisateur. Enfin, la méthode d'imputation de données proposée dans cette thèse pourrait être utilisée de manière opérationnelle dans un plus large éventail d'applications liées au suivi de la végétation (*e.g.*, pour le suivi de la production de la vigne[1]). Pour de telles applications, le nombre de parcelles disponibles pour l'analyse pourrait être un défi majeur.
- **Classification des parcelles anormales :** une perspective directement liée au point précédent est de pouvoir classifier les anomalies affectant les parcelles détectées. Une

[1] http://oenoview365.terranis.fr/

perspective à court terme est de fournir des outils pour aider l'utilisateur à identifier les différents types d'anomalies. Une perspective à long terme (et beaucoup plus difficile) est la classification automatique des parcelles anormales en différentes catégories. Une première tentative de classification supervisée a été faite dans León-López et al. (2021). Cependant, ces travaux mériteraient d'être approfondis pour être adaptés à un contexte opérationnel. Les principaux défis liés à cette tâche sont 1) le nombre relativement faible d'exemples d'anomalies annotées par des experts, 2) la diversité des différents types d'anomalies et 3) le fait que les parcelles anormales puissent appartenir à plusieurs catégories simultanément (par exemple, une parcelle anormales peut à la fois avoir une croissance tardive et un problème d'hétérogénéité). Enfin, il ne serait pas réaliste en pratique de devoir étiqueter chaque nouvel ensemble de données.

- **Types de cultures :** une perspective évidente consiste à appliquer la méthode proposée à d'autres types de cultures ou de végétation. De premiers tests ont été faits sur la vigne, avec des résultats prometteurs (ces résultats ne sont pas reportés dans ce manuscrits par souci de concision). Nous avons observé de bons résultats concernant la reconstruction de données manquantes à l'aide de GMM robuste, même avec des jeux de données de petite taille (erreur absolue moyenne de reconstruction du NDVI médian d'environ 0,018 dans une base de données contenant 1200 parcelles). Pour les vignobles, l'utilisation d'un GMM est particulièrement intéressante car ce type de végétation peut avoir une grande diversité de développements phénologiques au cours d'une même saison de croissance (par exemple, la surface de l'inter-rang peut être couverte d'herbe pour certaines parcelles). L'un des principaux défis liés à ce type de culture est la taille relativement petite des parcelles, ce qui peut poser problème avec la résolution spatiale des données Sentinel.

- **Indices caractérisant la végétation :** d'autres indicateurs caractérisant la végétation pourraient être étudiés et comparés à ceux présentés dans ce manuscrit. Parmi les exemples évidents, nous pouvons citer les indicateurs biophysiques extraits d'images multispectrales (comme par exemple l'indice de surface foliaire) ou les indicateurs extraits en utilisant la phase du signal radar pour les images RSO (images dites *single look complex (SLC)* en anglais). En fonction de l'application finale, le choix des indicateurs de végétation et des statistiques spatiales peut être un enjeu important, car il est directement corrélé au type d'anomalies détectées par la suite.

- **Analyse au niveau pixel:** une autre perspective pourrait être de travailler au niveau du pixel. Rappelons tout d'abord que travailler au niveau de la parcelle présente plusieurs avantages, en particulier 1) cela réduit les coûts de calcul et de stockage, ce qui est un enjeu important dans les applications basées sur la télédétection (Inglada et al., 2017) et 2) cela permet de d'utiliser des images RSO sans avoir besoin de traitements supplémentaires pour réduire le bruit de chatoiement (*speckle noise* en anglais), traitements demandant également beaucoup de puissance de calcul. Toutefois, travailler au niveau du pixel augmenterait le nombre d'échantillons à analyser, ce qui pourrait être intéressant pour lutter contre le fléau de la dimension (*curse of dimensionality* en anglais). Prendre une décision au niveau du pixel pourrait également être intéressant pour localiser spatialement les zones présentant un développement anormal au sein

d'une parcelle.

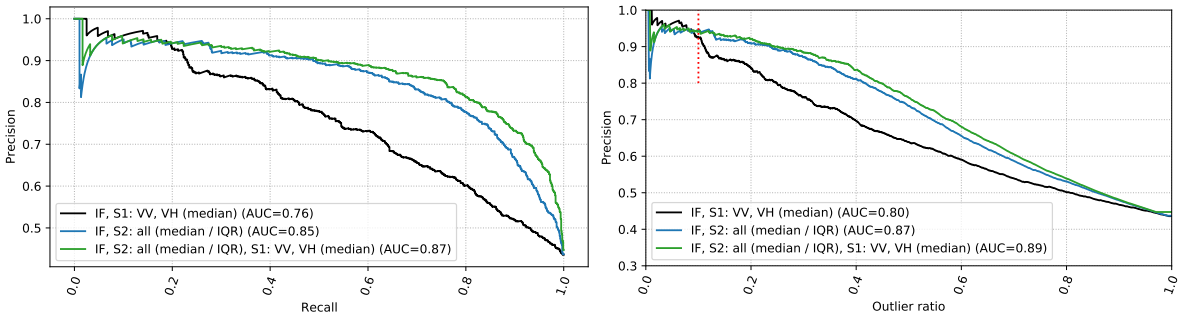
- **Analyse de plusieurs saisons de croissance :** une perspective intéressante à long terme est d'étendre le problème étudié dans ce manuscrit à l'analyse de plusieurs saisons de croissance. Divers défis devraient être relevés dans ce cadre là, comme par exemple le problème de la rotation des cultures, les problèmes causés par la disparité temporelle de l'acquisition des données satellites ou les défis liés à la variabilité interannuelle des cultures. Dans ce contexte, des recherches supplémentaires devraient être menées pour analyser l'intérêt des approches temporelles, par exemple pour analyser un petit nombre de parcelles provenant d'une nouvelle saison de culture à l'aide d'un modèle appris sur une saison précédente.

- **Régularisation des GMM :** nous avons constaté que la reconstruction des données manquantes avec des GMM est adaptée pour les indicateurs de végétation utilisés dans cette thèse. Cependant, nous pensons que ces résultats pourraient être encore améliorés en contraignant la structure des matrices de covariance, en particulier pour des données de grande dimension. Rappelons ici que plusieurs tentatives ont été faites sans succès au cours de cette thèse (par exemple en ajoutant une contrainte de parcimonie sur les matrices de précision ou de covariance). Une possibilité intéressante serait d'imposer une contrainte Toeplitz sur les matrices de covariance. Nous avons obtenu des résultats prometteurs sur des bases de données synthétiques. Cependant les expériences sur des données réelles n'ont pour le moment pas été concluantes, principalement en raison de la résolution temporelle réduite des données S2 qui casse la structure Toeplitz des matrices de covariance. L'utilisation des données SAR pourrait être un moyen de résoudre ce problème. De plus, toutes nos expérimentations ont été réalisées en utilisant une structure Toeplitz unique. L'utilisation de caractéristiques multiples nécessite un travail théorique plus poussé, par exemple en permettant la prise en compte de matrices de covariance Toeplitz en bloc.

Appendix A

A.1 Complementary information about precision vs. outlier ratio curves

It was noticed that precision vs. outlier ratio curves are better suited for our study than other similar evaluation curves (*e.g.*, precision vs. recall or ROC curves). Indeed, the outlier ratio is a parameter that is easier to adjust without ground-truth, by selecting the parcels with the highest outlier scores. Moreover, when analyzing these curves one can focus on realistic values of the outlier ratio (e.g., precision obtained when detecting more than 50% of the data instances seems not adapted to our problem). A comparison between precision vs. outlier ratio and precision vs. recall curves is displayed in Figure A.1, where it can be seen that both curves lead to the same conclusions (similar results are obtained with ROC curves). Evaluating outlier detection algorithm is difficult in practice, as pointed out in (Aggarwal, 2017, Chapter 1.7). An interesting discussion on such evaluation curves is available for instance in Saito and Rehmsmeier (2015).



(a) (b)

Figure A.1: (a) Precision vs. outlier ratio and (b) Precision vs. recall obtained using the IF algorithm on the rapeseed parcels for a complete growing season analysis.

Appendix B

B.1 Complementary results on various factors influencing the outlier detection results

This section provides results on various factors influencing the detection results, which are discussed in Chapter 3.

B.1.1 Effect of the outlier ratio

Three experiments were run using the median and IQR statistics derived from S2 images, the median statistics derived from S1 images and the IF algorithm, varying the outlier ratio in $\{0.1, 0.2, 0.3\}$. The percentages of detected parcels in the different anomaly categories for each of these experiments are depicted in Figure B.1. For an outlier ratio of 10%, the detected anomalies are mostly concentrated in wrong types, late growth and global heterogeneity which is relevant and confirms the observations made in the main document of this study. Moreover, for this outlier ratio, 45% of the detected parcels belong to the category referred to as "global heterogeneity", which is coherent since this type of anomaly is (generally) strongly affecting the crop development of the parcels. Increasing the outlier ratio allows anomalies affecting smaller time periods of the season to be detected, such as early flowering and senescence problems in accordance to the observation made during labeling. For an outlier ratio of 30%, much more false positives are detected (parcels labeled as normal). These results show that the IF algorithm provides a relevant anomaly score since more severe anomalies have higher anomaly scores. Moreover, because the score given by IF is computed only once, there is no need to run the algorithm several times when changing the outlier ratio and the amount of parcel to be detected can be easily adapted to the users' needs. Finally, for a generic analysis, choosing an outlier ratio of 20% is a good balance between the precision of the detection results and the amount of parcel to be detected.

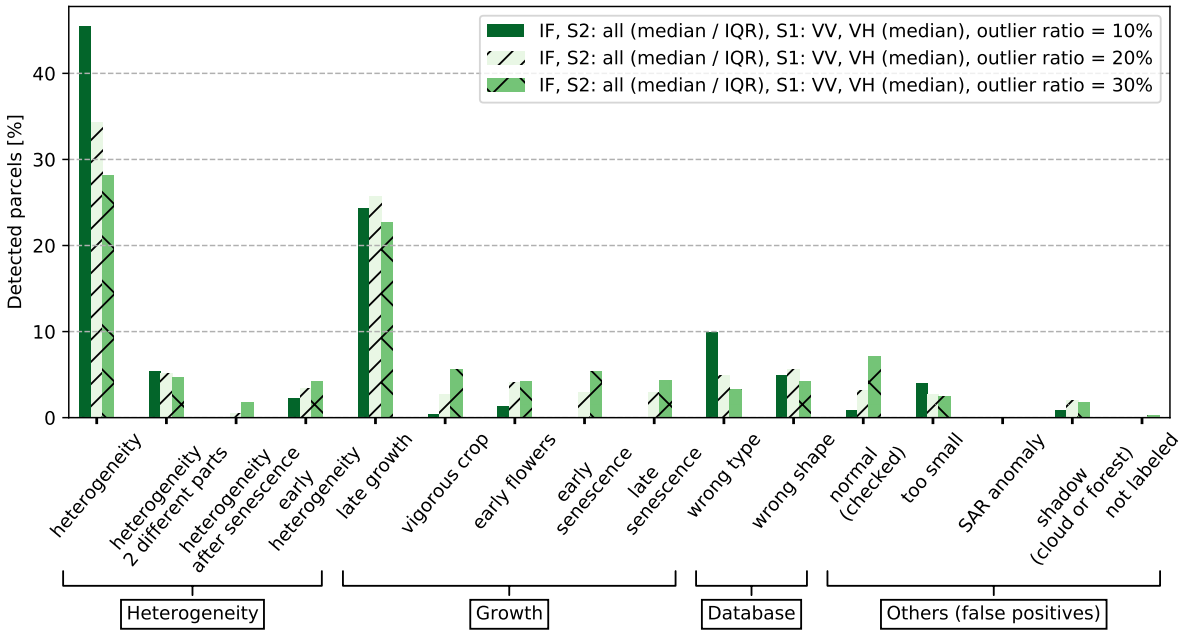


Figure B.1: 100×(Number of detected parcels in each category / Number of detected parcels). Various outlier ratio are tested with the same set of features and the IF algorithm for a complete growing season analysis (rapeseed crops).

B.1.2 Effect of adding new statistics

All the previous experiments were conducted using the median and IQR of S2 data as statistics computed at the parcel-level. This section investigates two new statistics for S2 data, namely the skewness and kurtosis (*i.e.*, the normalized third and fourth order moments of the features). Figure B.2 shows the precision vs. outlier ratio when using the IF algorithm and these two additional statistics computed from S2 images to detect anomalies in rapeseed parcels. All the parcels are labeled for outlier ratios that are at least smaller than 10% (less tests were made with skewness and kurtosis statistics as poor results were obtained). It can be observed in this figure that even for an outlier ratio lower than 5%, using skewness and kurtosis statistics leads to a significant difference in the precision results. One issue encountered when using these new statistics is the detection of too subtle anomalies that are not always related to agronomic anomalies. Using the median only is also tested but provides a lower average precision score. This analysis confirms the importance of IQR statistics, which allows a larger number of relevant anomalies to be detected, and in particular heterogeneity problems. This section showed that using median and IQR statistics of S2 features computed at the parcel level is recommended for crop monitoring.

Similarly, the effect of changing zonal statistics for S1 data was tested. More precisely, results obtained using 1) additional IQR statistics and 2) mean statistics instead of the median are provided in Figure B.3. It appears that results obtained using the mean are very similar to the one obtained using the median. However, adding IQR statistics significantly decrease the average precision since more false positives are detected.

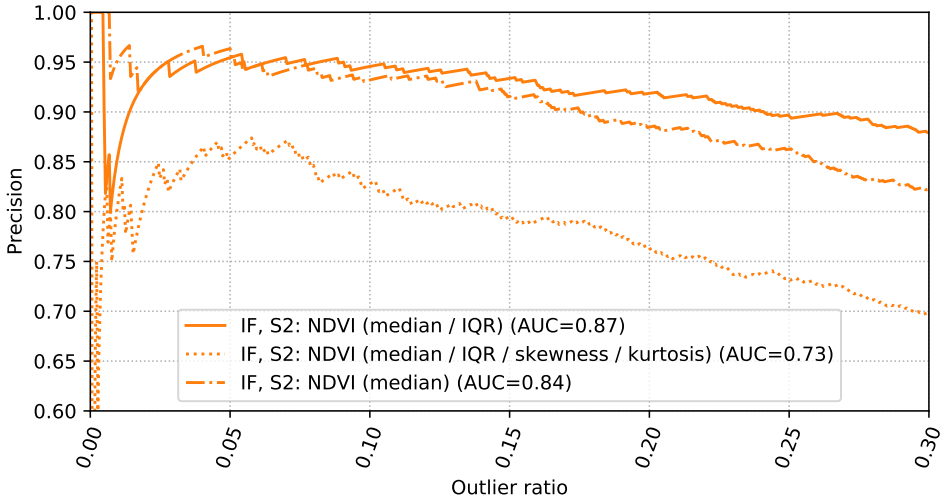


Figure B.2: Precision vs. outlier ratio for a complete growing season analysis of the rapeseed parcels. Various statistics of the NDVI are compared using the IF algorithm.

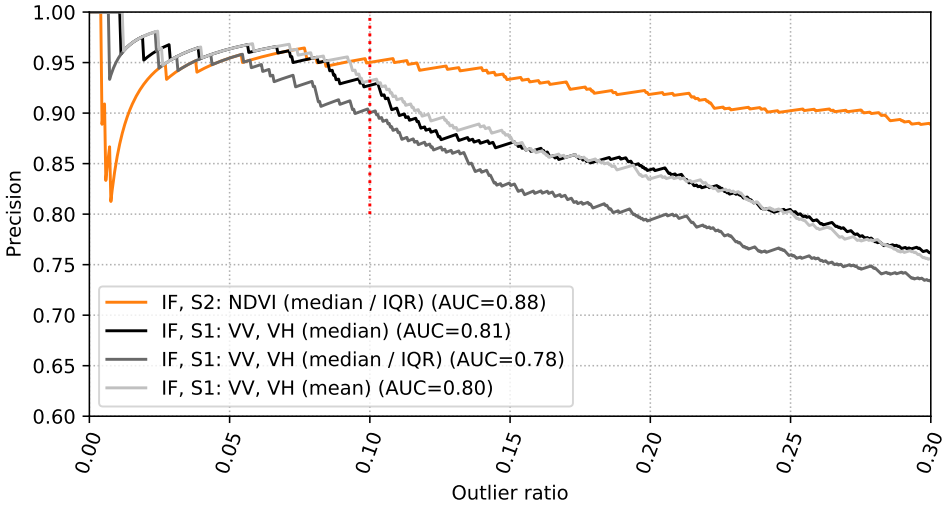


Figure B.3: Precision vs. outlier ratio for a complete growing season analysis of the rapeseed parcels. Various statistics of S1 back-scattering coefficients are compared using the IF algorithm.

B.1.3 Effect of missing S2 images

Two scenarios were investigated to evaluate the effect of missing S2 images.

- Scenario 1: the proposed approach was investigated using 6 S2 images instead of 13 to analyze the influence of a reduced amount of S2 images through the season. Only 1 image out of 2 was considered for the detection (the first S2 image was not used, the second S2 image was used and so on). Precision vs. outlier ratio curves are presented in Figure B.4, where it can be observed that the proposed method is robust to missing

S2 images.

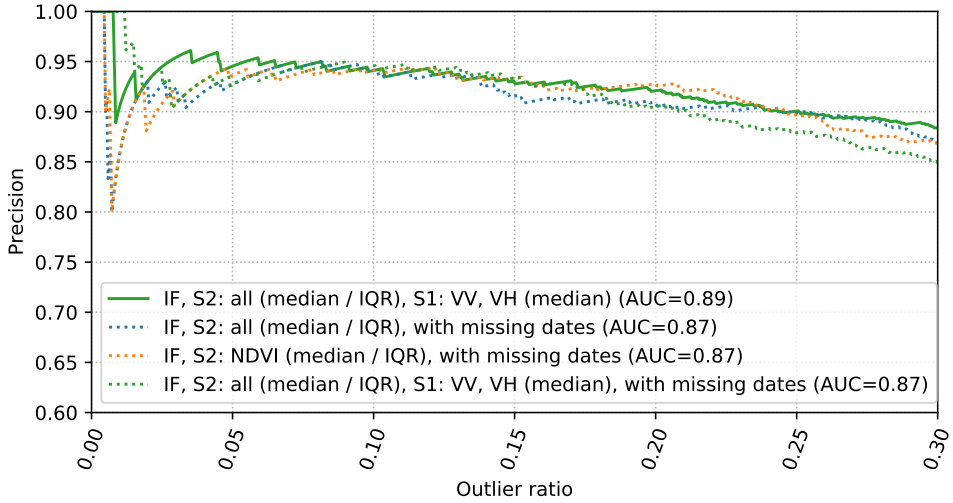


Figure B.4: Precision vs. outlier ratio for a complete season analysis of the rapeseed dataset. Missing dates means that only 1 S2 image out of 2 was taken (6 S2 images instead of 13).

- Scenario 2: another experiment was conducted to evaluate the effect of missing S2 images during the first part of the growing season (e.g, more clouds during winter). Precisely, we consider only 7 dates of S2 data between May and June that are used jointly with all S1 images. Precision vs. outlier ratio curves are presented in Figure B.5. In that case, using S1 images improve significantly the precision of the results. The reason is that using S1 features allows the algorithm to detect almost the same amount of late growth crops when compared to using a complete season of S2 images which is understandable since S1 data are well suited to detect growth anomalies. These results confirm the interest of using S1 features as a complement to S2 sparse time series.

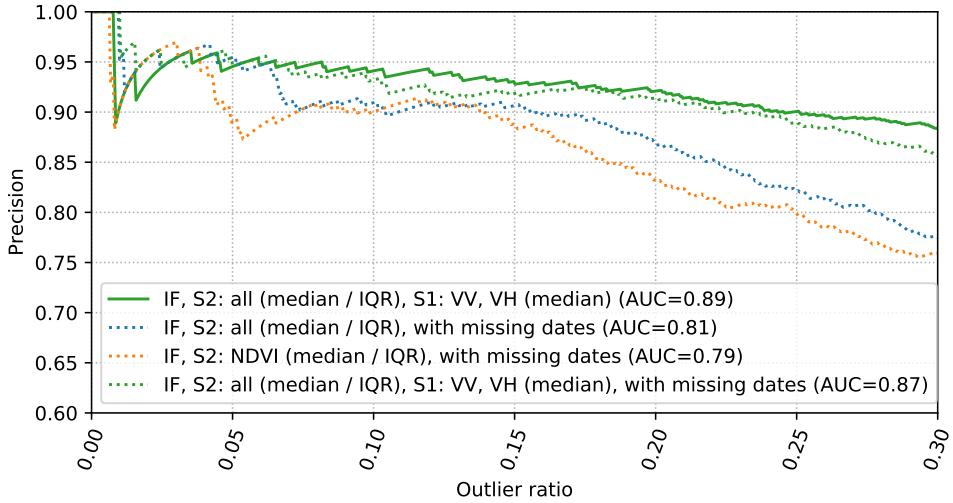


Figure B.5: Precision vs. outlier ratio for complete season analysis of the rapeseed dataset. Missing dates means that only the S2 images acquired after April were used (7 images).

B.1.4 Mid-season analysis

A mid-season analysis (using only dates before February) was conducted for multiple reasons detailed in Section 3.2. A first experiment was made with the best sets of features selected in Section 3.6 for a complete season analysis using rapeseed parcels. Results displayed in Figure B.6 show that even with a small number of images, many agronomic anomalies are detected (best precision=87.7% for an outlier ratio equal to 20%). This confirms the previous results found in the case of missing S2 images. Figure B.6 also shows that the best results are again obtained using all S1 and S2 features jointly with a higher average precision since more actual anomalies are detected for larger outlier ratios (*e.g.*, the precision is 5% better for an outlier ratio fixed to 30%).

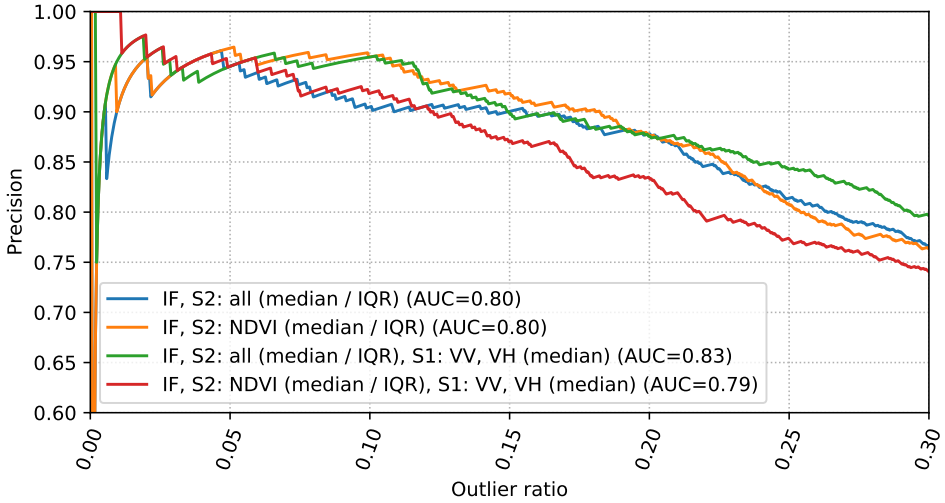


Figure B.6: Precision vs. outlier ratio for a mid-season analysis of rapeseed parcels (all images available before February). Various sets of features are compared using the IF algorithm.

The impact of a mid-season analysis regarding the different categories of detected anomalies is depicted in Figure B.7. In this case, almost no senescence problems are detected, which is easy to understand. Even with only 3 S2 images acquired between October and December, most other agronomic anomalies are detected by the algorithm. A mid-season analysis is able to detect more late growth anomalies and fewer heterogeneous parcels because late growth is impacting mostly the beginning of the season (especially for rapeseed crops). Finally, more false positives are detected with a mid-season analysis, which can be understood since the amount of potential anomalies to be detected is lower.

Complementary results for a mid season analysis are briefly presented in what follows since they confirm the observations made for a complete growing season analysis. The IF algorithm provides overall better results (AUC=0.83) and is more robust to changes. The AE performs slightly worse than IF (AUC=0.81), especially for outlier ratios greater than 20%. OCSV (AUC=0.79) and LoOP (AUC=0.77) perform significantly worse in this case. These differences in performance can be explained by the fact that the parameters of OC-SVM, LoOP and AE algorithms are more difficult to tune compared to the IF algorithm.

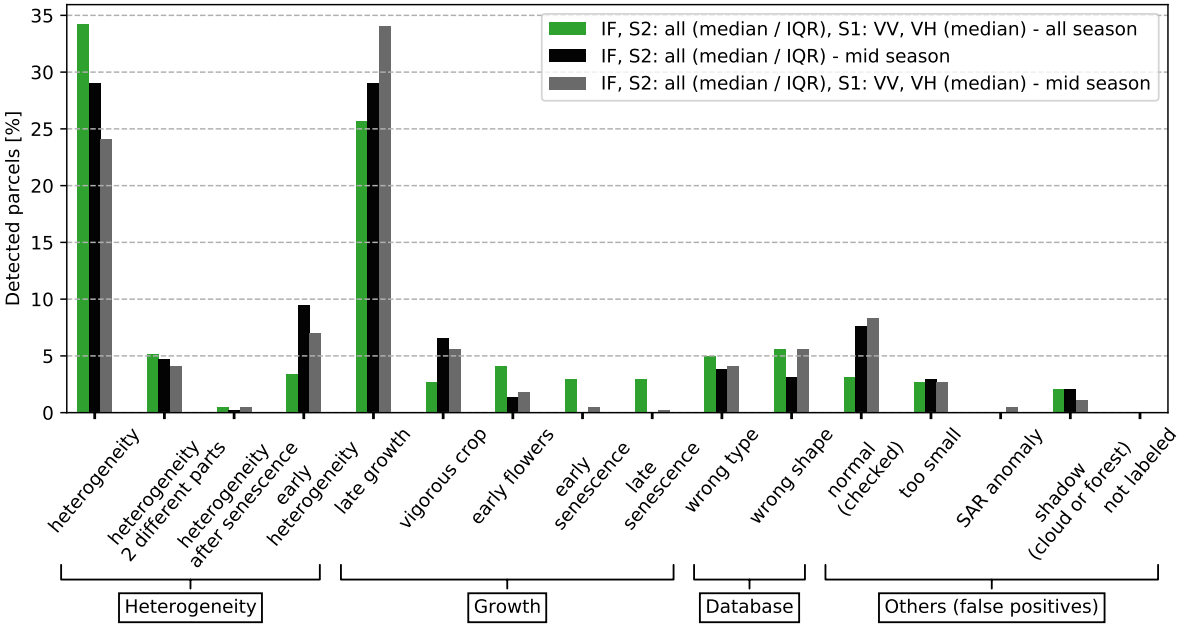


Figure B.7: 100×(Number of detected parcels in each category / Number of detected parcels). Results obtained for a mid season analysis (before February) and a complete growing season analysis are compared for a outlier ratio equal to 10% in the rapeseed dataset.

Regarding the influence of the outlier ratio, as for a complete season analysis increasing its value logically leads to detect more subtle anomalies (i.e., affecting a limited time interval) and more false positives, which confirms the relevance of the anomaly score given by IF. Almost no early heterogeneity and vigorous crop is detected with an outlier ratio of 10%. Early heterogeneity is a more subtle anomaly than global heterogeneity, which confirms separation between these two categories. Finally, when using S1 data only, the detection results obtained for an outlier ratio of 10% are still accurate with a precision equal to 89.6%. These results confirm that S1 images are adapted to an early season analysis, especially thanks to an easier detection of late growth problems.

B.1.5 Robustness to changes in parcel boundaries

The robustness of the proposed method to changes in the parcel boundaries was validated using another parcel delineation system for the rapeseed growing season. To that extent, 2118 parcel delineations resulting from the French Land Parcel Identification System (LPIS) was considered. The French LPIS is also known as *Registre Parcellaire Graphique* (RPG). This database is available with an open license [2] and is updated yearly (in general with a delay of 2 years) on the basis of the farmer's Common Agricultural Policy (CAP) (Barbottin et al., 2018). For comparison purposes, each parcel of database used in the main document was intersected with a corresponding LPIS parcel. Some parcels were not defined in the LPIS file, which explains why the number of parcels available for the LPIS analysis is slightly smaller than the number of parcels obtained when using the customer database.

Examples of parcel delinations obtained with LPIS and the proprietary parcellation system are depicted in Figure B.8. The parcel frontiers obtained using LPIS are generally less accurate than those resulting from the proprietary system motivating the use of a buffer around the different parcels and robust zonal statistics.

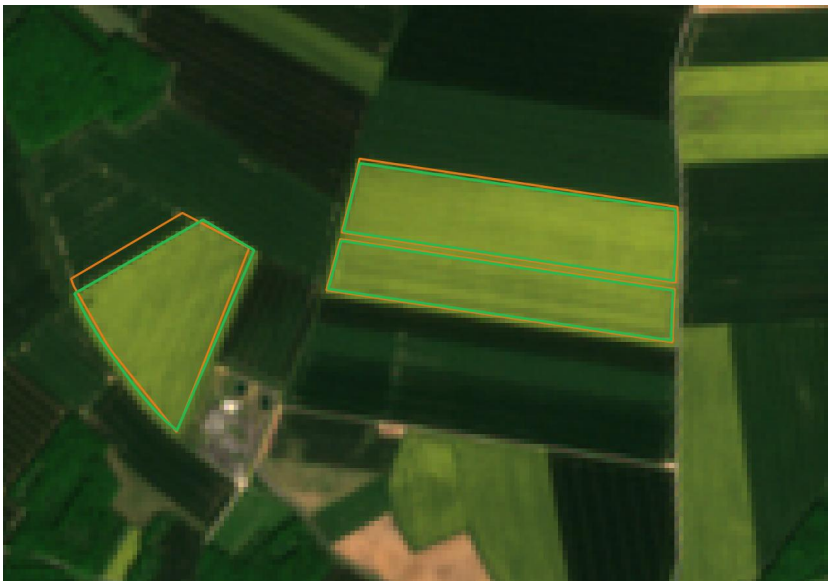


Figure B.8: Example of parcel boundaries (rapeseed crop, growing season 2017/2018). In orange: customer database, in green: LPIS database.

Anomaly detection was run with an outlier ratio of 20% using these two different databases. The numbers of detected anomalies for each category are depicted in Figure B.9. No significant difference can be observed when using the customer and LPIS parcels, showing that the proposed detection method is robust to this type of changes (probably because robust zonal statistics are used for anomaly detection).

[2] https://www.data.gouv.fr/fr/datasets/58d8d8a0c751df17537c66be/, online accessed 8 July 2020

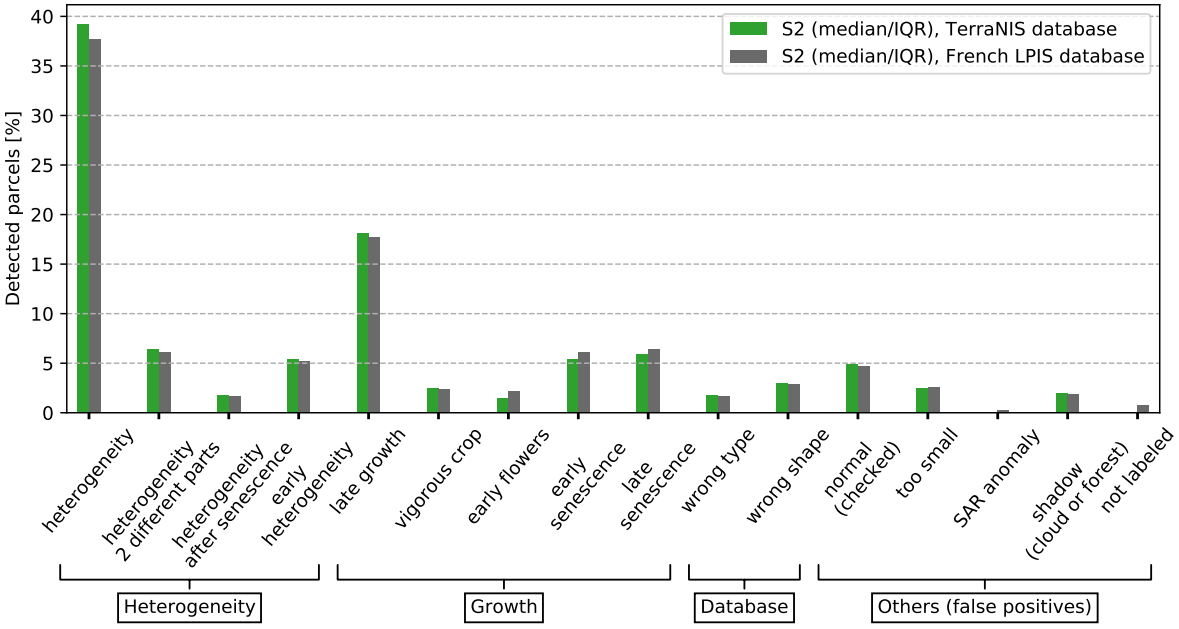


Figure B.9: 100×(Number of detected parcels in each category / Number of detected parcels). LPIS and proprietary parcellation databases are compared with the IF algorithm and an outlier ratio equal to 20%.

Appendix C

This appendix provides complementary experiments conducted on rapeseed and wheat crops regarding the reconstruction of vegetation indices computed using Sentinel-2 images, which are presented in Chapter 4.

C.1 Examples of data imputation for rapeseed and wheat parcels

C.1.1 Rapeseed parcel

To have an easier appreciation of the challenges related to the data imputation in S2 data, examples of reconstruction of 4 different time series (median NDVI (a), IQR NDVI (b), median NDWI (c) and IQR NDWI (d)) for a specific parcel are displayed in Figure C.1. For that experiment, the first 4 S2 acquisitions have missing values and 50% of the parcels are affected, making the imputation problem particularly difficult.

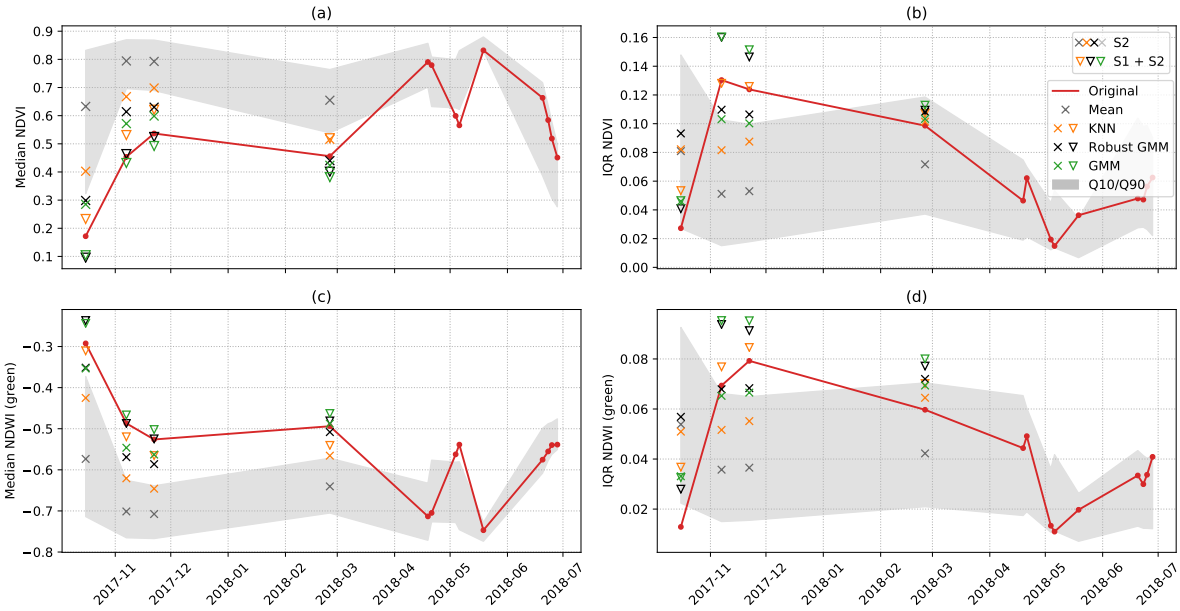


Figure C.1: For a specific rapeseed parcel, imputation of (a) median NDVI, (b) IQR NDVI, (c) median NDWI (green) and (d) IQR NDWI (green). The crosses correspond to imputations obtained by using S2 images only whereas triangles correspond to the joint use of S1 and S2 features. The gray area is filled between the 10th and 90th percentiles values of the whole dataset. 50% of the parcels are affected by missing values.

It can be observed that the analyzed parcel (red curve) has a late development during the first part of the growing season, which is an atypical behavior making the reconstruction

task even harder. Here, the mean imputation is displayed (gray crosses) to show that this method can be problematic if the values to be imputed are unusual. Overall, the robust GMM (black) is generally more accurate, with almost perfect reconstruction for the median features in November. The reconstruction of the IQR values tends to be overestimated in that case, even if the heterogeneous behavior is captured. In this extreme example, the interest of using S1 data (triangles) is particularly visible: since no data is available during the first part of the growing season, using this information is of crucial importance for all the tested methods. Finally, one can observe that the KNN imputation can provide good results (especially when using S1 data), but is generally less accurate (*e.g.*, median NDVI is constantly underestimated).

C.1.2 Wheat parcel

Figure C.2 provides an example similar to the one displayed in Figure C.1 but for a wheat parcel with a late and heterogeneous development. One can notice that the robust GMM imputations are close to the original values and that again using S1 data is helpful to capture the low vigor of the crop.

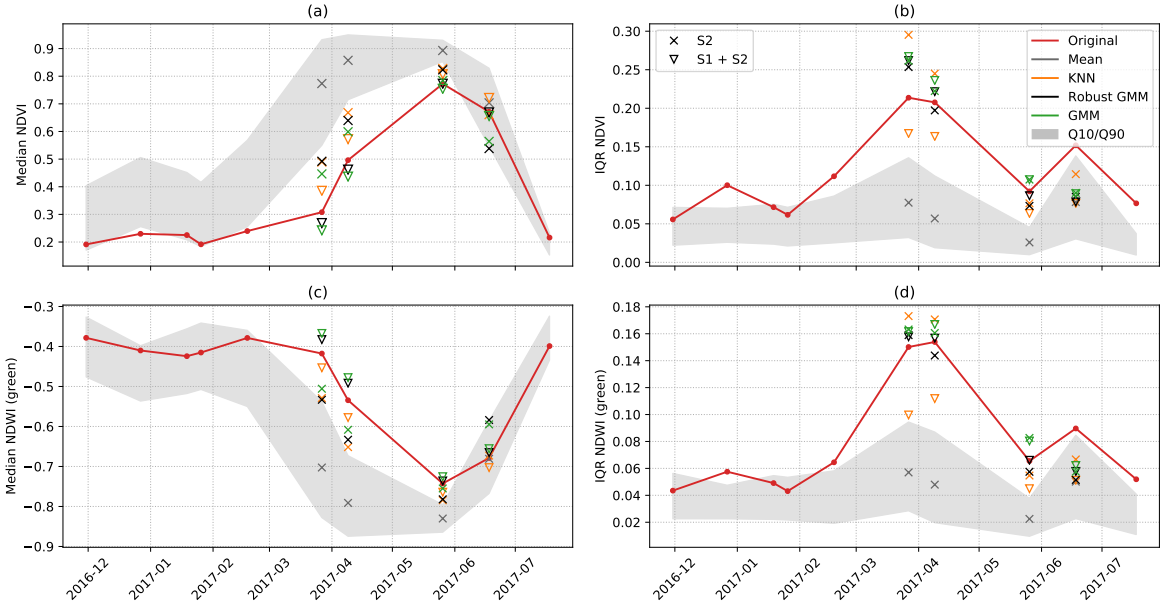


Figure C.2: For a specific rapeseed parcel, imputation of (a) median NDVI, (b) IQR NDVI, (c) median NDWI (green) and (d) IQR NDWI (green). The crosses correspond to imputations obtained by using S2 images only whereas triangles correspond to the joint use of S1 and S2 features. The gray area is filled between the 10th and 90th percentiles values of the whole dataset. 50% of the parcels are affected by missing values.

C.2 Day by day imputation

C.2.1 Rapeseed crops

An experiment was conducted by removing S2 features at each S2 acquisition (50% of the crop parcels are affected). The results obtained for the normalized S2 features are depicted in Figure C.3 (similar results have been observed when looking at specific features and are not plotted here for conciseness). In this example, two dates are more difficult to impute: the first acquisition (during sowing) and the data acquired at the end of February during winter. On the other hand, some stages of the growing season can be reconstructed with significantly lower MAE. Again, GMM imputation outperforms KNN imputation. Moreover, using S1 data is also useful to improve the results (particularly for the most difficult imputations).

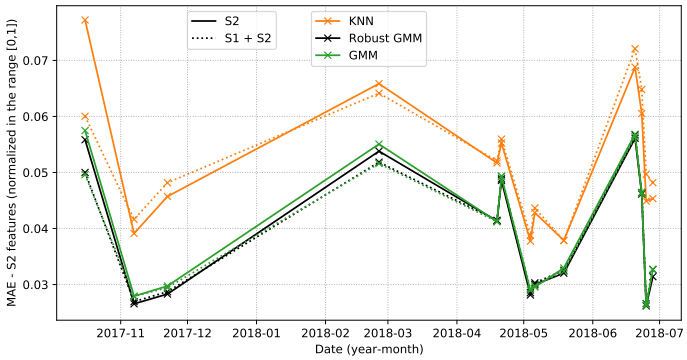


Figure C.3: Rapeseed crops are analyzed. X-axis: S2 acquisition with missing values: for each acquisition with missing data, 50% of the parcels are affected. Y-axis: MAE of the normalized S2 features the parcels. The solid lines results are obtained using only the S2 data, whereas dashed lines were obtained using both the S1 and S2 data. Results are averaged after 50 iterations

C.2.2 Wheat crops

The experiments presented in Figure C.3 were also conducted for wheat crops. Results are displayed in Figure C.4 for the normalized S2 signal. The robust GMM always provides the best reconstructions, in particular for the most difficult imputations. The features extracted from the S2 images acquired in June are the most difficult to reconstruct. For this acquisition, using S1 data improves the imputation significantly. The other experiments conducted on rapeseed crops lead to the same conclusions for the wheat parcels (not shown here).

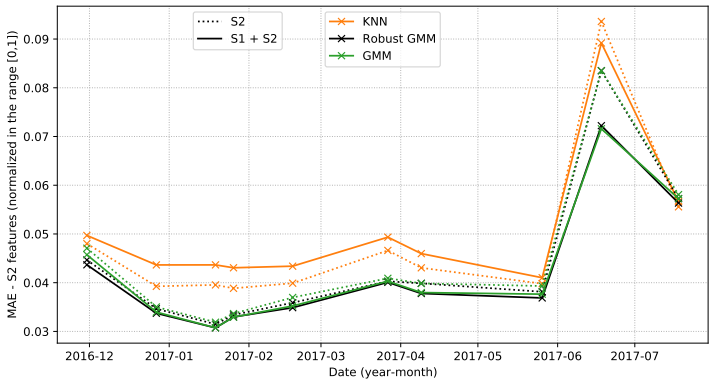


Figure C.4: Analysis of wheat crops. X-axis: S2 acquisitions with missing values: for each acquisition with missing data, 50% of the parcels are affected. Y-axis: MAE for normalized S2 features (all the S2 indicators are considered). The solid lines are obtained using S2 data only, whereas the dashed lines were obtained using both S1 and S2 data. Results are averaged using 50 Monte Carlo runs, each time 50% of the parcels are affected by missing data.

C.3 Detailed results for a specific S2 acquisition with missing data

The MAE, RMSE and coefficient of determination (R^2) were computed by averaging the results obtained after 500 simulations. For each simulation, all the S1 and S2 data were used and 50% of the parcels have missing values at a same random S2 acquisition. The obtained scores for rapeseed crops are provided in Table C.1. These results confirm that the robust GMM provides the best reconstructions overall (it is also the case when compared to the classical GMM). Some features are harder to reconstruct: it is particularly the case for the MCARI/OSAVI statistics, and more generally for the IQR of the different VI. IQR is not well captured by the S1 data and has less smooth time variations, which could explain this results.

The same experiment was done for wheat crops whose results are reported in Table C.2.

Table C.1: Regression scores (MAE, RMSE, R^2) obtained on the rapeseed dataset (200 MC simulations). Each time, 50% of the parcels have missing values at one S2 acquisition. S1 and S2 features are used to impute missing values. Standard deviation (std) is added in parenthesis. KNN and Robust-GMM (R-GMM) are compared, best results are in bold.

	MAE (std)		RMSE (std)		R^2 (std)	
Feature / Algorithm	KNN	R-GMM	KNN	R-GMM	KNN	R-GMM
median(NDVI)	0.029 (0.008)	**0.013 (0.007)**	0.042 (0.009)	**0.021 (0.009)**	0.71 (0.17)	**0.92 (0.09)**
median($NDWI_{GREEN}$)	0.021 (0.007)	**0.010 (0.005)**	0.030 (0.007)	**0.016 (0.006)**	0.64 (0.15)	**0.88 (0.10)**
median($NDWI_{SWIR}$)	0.029 (0.008)	**0.012 (0.007)**	0.043 (0.007)	**0.019 (0.009)**	0.77 (0.14)	**0.95 (0.04)**
median(GRVI)	0.027 (0.007)	**0.014 (0.007)**	0.037 (0.008)	**0.020 (0.009)**	0.72 (0.16)	**0.89 (0.10)**
median(MCARI/OSAVI)	133 (55)	**79 (46)**	180 (66)	**117 (64)**	0.61 (0.23)	**0.81 (0.16)**
IQR(NDVI)	0.015 (0.005)	**0.008 (0.005)**	0.024 (0.008)	**0.014 (0.007)**	0.58 (0.10)	**0.83 (0.13)**
IQR($NDWI_{GREEN}$)	0.009 (0.003)	**0.006 (0.003)**	0.016 (0.004)	**0.010 (0.005)**	0.52 (0.08)	**0.80 (0.14)**
IQR($NDWI_{SWIR}$)	0.018 (0.003)	**0.009 (0.005)**	0.028 (0.005)	**0.014 (0.007)**	0.57 (0.12)	**0.86 (0.14)**
IQR(GRVI)	0.013 (0.004)	**0.008 (0.004)**	0.019 (0.005)	**0.13 (0.006)**	0.51 (0.15)	**0.76 (0.18)**
IQR(MCARI/OSAVI)	62 (22)	**42 (20)**	93 (30)	**65 (30)**	0.40 (0.18)	**0.68 (0.21)**

Table C.2: Regression scores (MAE, RMSE, R^2) obtained on the wheat dataset (200 MC simulations). Each time, 50% of the parcels have missing values at one S2 acquisition. S1 and S2 features are used to impute missing values. Standard deviation (std) is added in parenthesis. KNN and Robust-GMM (R-GMM) are compared, best results are in bold.

	MAE (std)		RMSE (std)		R^2 (std)	
Feature / Algorithm	KNN	R-GMM	KNN	R-GMM	KNN	R-GMM
median(NDVI)	0.032 (0.013)	**0.020 (0.010)**	0.044 (0.015)	**0.029 (0.013)**	0.70 (0.27)	**0.81 (0.26)**
median($NDWI_{GREEN}$)	0.026 (0.008)	**0.018 (0.008)**	0.035 (0.009)	**0.024 (0.006)**	0.67 (0.30)	**0.79 (0.29)**
median($NDWI_{SWIR}$)	0.031 (0.012)	**0.020 (0.009)**	0.042 (0.015)	**0.029 (0.012)**	0.69 (0.21)	**0.82 (0.21)**
median(GRVI)	0.032 (0.020)	**0.024 (0.018)**	0.042 (0.027)	**0.033 (0.024)**	0.63 (0.18)	**0.75 (0.27)**
median(MCARI/OSAVI)	127 (93)	**106 (113)**	206 (215)	**178 (228)**	0.51 (0.25)	**0.64 (0.28)**
IQR(NDVI)	0.015 (0.008)	**0.010 (0.007)**	0.024 (0.009)	**0.017 (0.01)**	0.41 (0.18)	**0.68 (0.26)**
IQR($NDWI_{GREEN}$)	0.011 (0.004)	**0.008 (0.004)**	0.018 (0.006)	**0.012 (0.005)**	0.32 (0.19)	**0.59 (0.27)**
IQR($NDWI_{SWIR}$)	0.015 (0.007)	**0.010 (0.005)**	0.022 (0.008)	**0.016 (0.008)**	0.29 (0.14)	**0.60 (0.23)**
IQR(GRVI)	0.014 (0.011)	**0.012 (0.010)**	0.021 (0.014)	**0.18 (0.013)**	0.35 (0.20)	**0.55 (0.27)**
IQR(MCARI/OSAVI)	**83 (82)**	88 (117)	163 (210)	**154 (225)**	0.29 (0.20)	**0.44 (0.28)**

C.4 Imputation results using the MICE algorithm

The experiment presented in Figure 6 of the main document was conducted using various imputation algorithms. Results obtained after adding the Multiple Imputation by Chained Equations (MICE) algorithm are displayed in Figure C.5 (using S1 and S2 data jointly). Overall, the MICE algorithm provides imputation slightly better than the KNN algorithms (except when the amount of missing S2 images is greater than 50%). In any cases, using GMM imputation is significantly better.

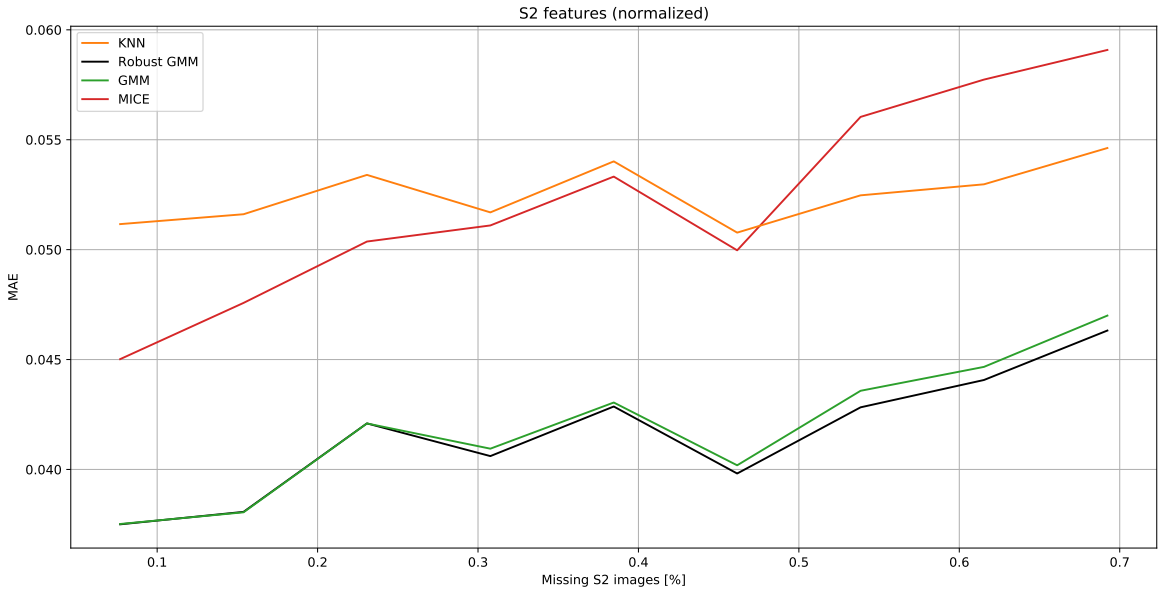


Figure C.5: Analysis of rapeseed crops. X-axis: percentage of S2 with missing values. Y-axis: MAE for the normalized S2 features (all the S2 indicators are considered). Results are obtained using S1 and S2 data jointly. The results are averaged using 50 MC runs, each time 50% of the parcels are affected by missing data for each S2 image with missing data.

C.5 Detecting anomalies in the rapeseed dataset with additional S2 images

Considering the 2218 rapeseed parcels analyzed in Chapter 3 as baseline, we added 8 S2 images to the database, yielding a total of 21 S2 images. The new images are cloudy, as shown in Figure C.6, which explains why they were not selected in the analysis conducted in Chapter 3.

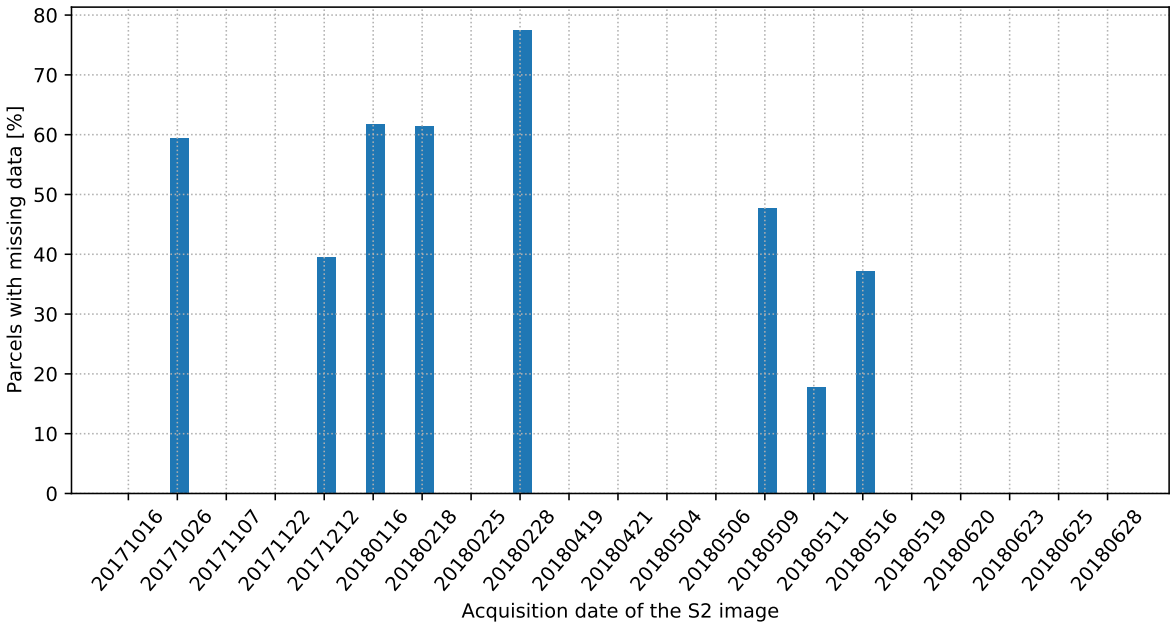


Figure C.6: Percentage of rapeseed parcels with missing data for each S2 acquisition. For 13 acquisitions, the 2218 parcel analyzed have no missing data.

Precision vs. outlier ratio curves have been computed for the two different rapeseed datasets (i.e., the initial dataset with 13 S2 images, and the extended dataset with 21 S2 images) and are displayed in Figure C.7. One can see that similar precisions are obtained with the two datasets independently from the imputation method, confirming the robustness of the detection method with respect to changes in the features. Overall, using more S2 images leads to a precision slightly higher, especially when detecting more subtle anomalies (i.e., for an outlier ratio close to 20%). Using the new S2 images, we have been able to detect new anomalies associated with the analyzed parcels (these changes are only considered when using the additional S2 images). Finally, some false positive have been detected because of errors in the cloud masks. We have observed that these errors tend to occur more often when adding very cloudy images.

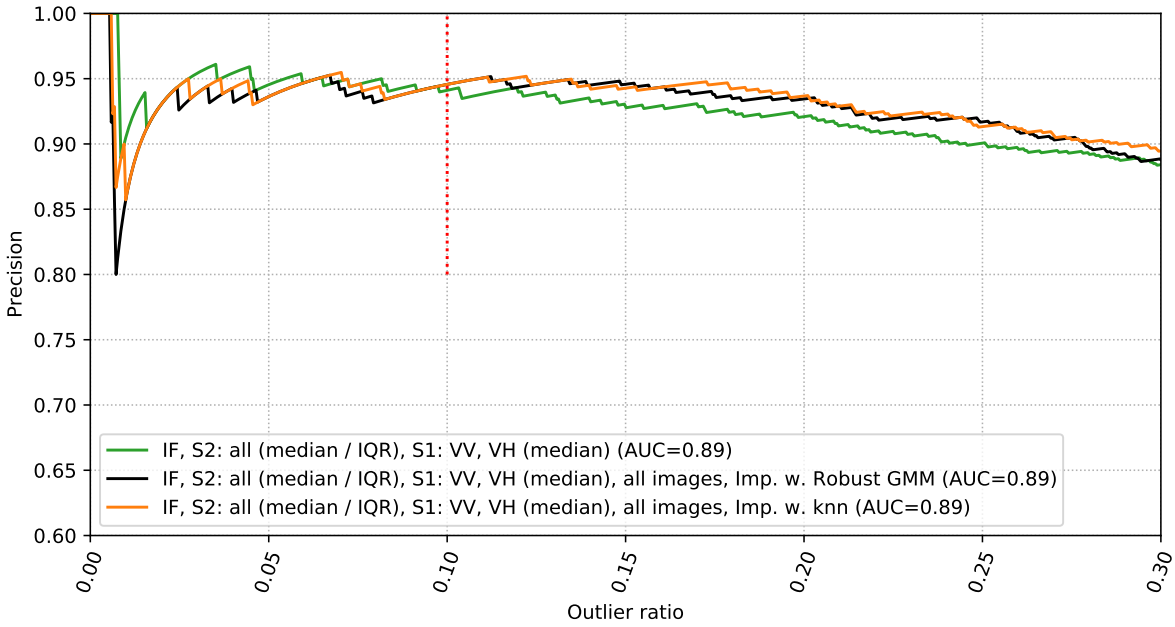


Figure C.7: Precision vs. outlier ratio when using the IF algorithm on the rapeseed parcels. Green: original dataset analyzed in Chapter 3, black: dataset extended with new S2 images and imputed with Robust GMM, orange: dataset extended with new S2 images and imputed with KNN.

Bibliography

Abdikan, S., Balik Sanli, F., Üstüner, M., Calò, F., 2016. Land cover mapping using Sentinel-1 SAR data, in: Proc. ISPRS, Prague, Czech Republic. pp. 757–761. doi:10.5194/isprs-archives-XLI-B7-757-2016.

Aggarwal, C.C., 2017. Outlier Analysis. 2nd ed., Springer International Publishing, Cham. doi:10.1007/978-3-319-47578-3_3.

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.

Areal, F.J., Jones, P.J., Mortimer, S.R., Wilson, P., 2018. Measuring sustainable intensification: Combining composite indicators and efficiency analysis to account for positive externalities in cereal production. Land Use Policy 75, 314–326. doi:https://doi.org/10.1016/j.landusepol.2018.04.001.

Atzberger, C., 2013. Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. Remote Sens. 5, 949–981. doi:10.3390/rs5020949.

Atzberger, C., Eilers, P.H.C., 2011a. Evaluating the effectiveness of smoothing algorithms in the absence of ground reference measurements. Int. J. Remote Sens. 32, 3689–3709.

Atzberger, C., Eilers, P.H.C., 2011b. A time series for monitoring vegetation activity and phenology at 10-daily time steps covering large parts of South America. Int. J. Digit. Earth 4, 365–386.

Bannari, A., Morin, D., Bonn, F., Huete, A.R., 1995. A review of vegetation indices. Remote Sens. Rev. 13, 95–120. doi:10.1080/02757259509532298.

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.

Baret, F., Houlès, V., Guérif, M., 2007. Quantification of plant stress using remote sensing observations and crop models: the case of nitrogen management. J. Exp. Bot. 58, 869–880. URL: http://www.jstor.org/stable/24036504.

Baum, L.E., Eagon, J.A., 1967. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bull. Am. Math. Soc. 73, 360 – 363. doi:bams/1183528841.

Baum, L.E., Petrie, T., 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. Ann. Math. Stat. 37, 1554 – 1563. URL: https://doi.org/10.1214/aoms/1177699147, doi:10.1214/aoms/1177699147.

Beck, P.S., Atzberger, C., Høgda, K.A., Johansen, B., Skidmore, A.K., 2006. Improved monitoring of vegetation dynamics at very high latitudes: A new method using MODIS NDVI. Remote Sens. Environ. 100, 321–334.

Betbeder, J., Rémy, F., Philippets, Y., Ferro-Famil, L., Baup, F., 2016. Contribution of multitemporal polarimetric synthetic aperture radar data for monitoring winter wheat and rapeseed crops. J. Appl. Remote Sens. 10, 026020. doi:10.1117/1.JRS.10.026020.

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

Borg, I., Groenen, P., 1997. Modern Multidimensional Scaling. Springer-Verlag New York.

Bouvet, A., Le Toan, T., Lam-Dao, N., 2009. Monitoring of the rice cropping system in the mekong delta using envisat/asar dual polarization data. IEEE Trans. Geosci. Remote Sens. 47, 517–526. doi:10.1109/TGRS.2008.2007963.

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.

Breunig, M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. LOF: Identifying density-based local outliers, in: Proc. ACM SIGMOD, Dallas, TX, USA. pp. 93–104.

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.

Calera, A., Campos, I., Osann, A., D'Urso, G., Menenti, M., 2017. Remote sensing for crop water management: From et modelling to services for the end users. Sensors 17, 1104. doi:10.3390/s17051104.

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

Campos-Taberner, García-Haro, Martínez, Sánchez-Ruiz, Gilabert, 2019. A copernicus Sentinel-1 and Sentinel-2 classification framework for the 2020+ european common agricultural policy: A case study in València (spain). Agronomy 9, 556. URL: http://dx.doi.org/10.3390/agronomy9090556, doi:10.3390/agronomy9090556.

Chandola, V., Banerjee, A., Kumar, V., 2009. Survey of anomaly detection. ACM Comput. Surveys 41, 15:1–15:58.

Constantinou, V., 2018. PyNomaly: Anomaly detection using local outlier probabilities (LoOP). J. Open Source Softw. 3, 845. doi:10.21105/joss.00845.

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

Das, S., Wong, W.K., Dietterich, T., Fern, A., Emmott, A., 2016. Incorporating expert feedback into active anomaly discovery, in: Proc. IEEE ICDM), Barcelona, Spain. pp. 853–858. doi:10.1109/ICDM.2016.0102.

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.

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.

Djamai, N., Fernandes, R., Weiss, M., McNairn, H., Goïta, K., 2019. Validation of the Sentinel simplified Level 2 product prototype processor (SL2P) for mapping cropland biophysical variables using Sentinel-2/MSI and Landsat-8/OLI data. Remote Sens. Environ. 225, 416 – 430. URL: http://www.sciencedirect.com/science/article/pii/S0034425719301117, doi:https://doi.org/10.1016/j.rse.2019.03.020.

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.

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.

van Emmerik, T., 2017. Water stress detection using radar. Ph.D. thesis. Delft University of Technology. Delft, Netherlands. doi:https://doi.org/10.4233/uuid:46f0b6e6-5592-4b05-983b-a04c8f0f88a8.

FAO, 2017. The future of food and agriculture - Trends and challenges. Food and Agriculture Organization of the United Nations, Rome URL: http://www.fao.org/3/i6583e/i6583e.pdf.

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.

Filipponi, F., 2019. Sentinel-1 GRD preprocessing workflow, in: Proc. ECRS-3, MDPI AG. p. 11. doi:10.3390/ecrs-3-06201.

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.

Gomiero, T., Pimentel, D., Paoletti, M.G., 2011. Environmental impact of different agricultural management practices: Conventional vs. organic agriculture. Crit. Rev. Plant. Sci. 30, 95–124. doi:10.1080/07352689.2011.554355.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z., 2019. XAI —explainable artificial intelligence. Sci. Robot. 4. doi:10.1126/scirobotics.aay7120.

Gómez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: A review. ISPRS J. Photogramm. Remote Sens. 116, 55 – 72. URL: http://www.sciencedirect.com/science/article/pii/S0924271616000769, doi:https://doi.org/10.1016/j.isprsjprs.2016.03.008.

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.

Hallac, D., Vare, S., Boyd, S., Leskovec, J., 2017. Toeplitz inverse covariance-based clustering of multivariate time series data, in: Proc. ACM KDD, Association for Computing Machinery, New York, NY, USA. p. 215–223. URL: https://doi.org/10.1145/3097983.3098060, doi:10.1145/3097983.3098060.

Hamdi, A., Frigui, H., 2015. Ensemble hidden Markov models with application to landmine detection. EURASIP J. Adv. Signal Process. 2015. doi:10.1109/36.298019.

He, Z., Li, S., Wang, Y., Dai, L., Lin, S., 2018. Monitoring rice phenology based on backscattering characteristics of multi-temporal RADARSAT-2 datasets. Remote Sens. 10, 340. doi:10.3390/rs10020340.

Hird, J.N., McDermid, G.J., 2009. Noise reduction of NDVI time series: An empirical comparison of selected techniques. Remote Sens. Environ. 113, 248–258.

Huber, P.J., 2011. Robust Statistics. Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 1248–1251. URL: https://doi.org/10.1007/978-3-642-04898-2_594, doi:10.1007/978-3-642-04898-2_594.

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.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017. Operational high resolution land cover map production at the country scale using satellite image time series. Remote Sens. 9, 95. URL: http://dx.doi.org/10.3390/rs9010095, doi:10.3390/rs9010095.

Jaakkola, T., Diekhans, M., Haussler, D., 1999. Using the Fisher kernel method to detect remote protein homologies, in: ISMB, AAAI Press, Menio Park, CA, USA. pp. 149–158.

Jolliffe, I.T., 1986. Principal Component Analysis. Springer New York, New York, NY. doi:10.1007/978-1-4757-1904-8_8.

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.

Keogh, E., Lin, J., Fu, A., 2005. HOT SAX: efficiently finding the most unusual time series subsequence, in: Proc. IEEE ICDM), Huston, TX, USA. pp. 8 pp.–. doi:10.1109/ICDM.2005.79.

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.

Klisch, A., Atzberger, C., 2016. Operational drought monitoring in Kenya using MODIS NDVI time series. Remote Sens. 8, 267.

Kramer, M.A., 1991. Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal 37, 233–243. doi:10.1002/aic.690370209.

Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A., 2009. LoOP: Local outlier probabilities, in: Proc. CIKM, Hong Kong, China. pp. 1649–1652. doi:10.1145/1645953.1646195.

Kumar, D., Rao, S., Sharma, J., 2013. Radar vegetation index as an alternative to NDVI for monitoring of soyabean and cotton, in: Proc. INCA, Jodhpur, India. pp. 91–96.

Kussul, N., Mykola, L., Shelestov, A., Skakun, S., 2018. Crop inventory at regional scale in Ukraine: developing in season and end of season crop maps with multi-temporal optical and SAR satellite imagery. Eur. J. Remote Sens. 51, 627–636. doi:10.1080/22797254.2018.1454265.

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.

Learning, L.B., 2001. Random forests. Mach. Learn. 45, 5–32. doi:https://doi.org/10.1023/A:1010933404324.

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. IEEE Trans. Geosci. Remote Sens. , 1–11doi:10.1109/TGRS.2021.3101127.

Liu, C., Chen, Z., Shao, Y., Chen, J., Hasi, T., Pan, H., 2019. Research advances of SAR remote sensing for agriculture applications: A review. J. Integr. Agric. 18, 506 – 525. URL: http://www.sciencedirect.com/science/article/pii/S2095311918620167, doi:https://doi.org/10.1016/S2095-3119(18)62016-7.

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.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67. doi:10.1038/s42256-019-0138-9.

López-Granados, F., 2011. Weed detection for site-specific weed management: mapping and real-time approaches. Weed Res. 51, 1–11. doi:https://doi.org/10.1111/j.1365-3180.2010.00829.x.

Mahlein, A.K., 2016. Plant disease detection by imaging sensors – parallels and specific demands for precision agriculture and plant phenotyping. Plant Disease 100, 241–251. doi:10.1094/PDIS-03-15-0340-FE.

Marzahn, P., Wegmuller, U., Mattia, F., Ludwig, R., 2012. "flashing fields" and the impact of soil surface roughness, in: Proc. IEEE IGARSS, Munich, Germany. pp. 6963–6966. doi:10.1109/IGARSS.2012.6351968.

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.

McNairn, H., Shang, J., 2016. A review of multitemporal synthetic aperture radar (SAR) for crop monitoring, in: Ban, Y. (Ed.), Multitemporal Remote Sensing: Methods and Applications. Springer International Publishing, Cham, Switzerland. chapter 15, pp. 317–340. doi:10.1007/978-3-319-47037-5_15.

Meroni, M., d'Andrimont, R., Vrieling, A., Fasbender, D., Lemoine, G., Rembold, F., Seguini, L., Verhegghen, A., 2021. Comparing land surface phenology of major european crops as derived from SAR and multispectral data of Sentinel-1 and -2. Remote Sens. Environ. 253, 112232. doi:https://doi.org/10.1016/j.rse.2020.112232.

Meroni, M., Fasbender, D., Rembold, F., Atzberger, C., Klisch, A., 2019. Near real-time vegetation anomaly detection with MODIS NDVI: Timeliness vs. accuracy and effect of anomaly computation options. Remote Sens. Environ. 221, 508–521.

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., 2021a. 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.

Mouret, F., Albughdadi, M., Duthoit, S., Kouamé, D., Rieu, G., Tourneret, J.Y., 2021b. Reconstruction of sentinel-2 derived time series using robust Gaussian mixture models —application to the detection of anomalous crop development. Under review .

Nasirzadehdizaji, R., Balik Sanli, F., Abdikan, S., Cakir, Z., Sekertekin, A., Ustuner, M., 2019. Sensitivity analysis of multi-temporal Sentinel-1 SAR parameters to crop height and canopy coverage. Appl. Sci. 9, 655. URL: http://dx.doi.org/10.3390/app9040655, doi:10.3390/app9040655.

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.

Newbold, T., Hudson, L., Hill, S., Contu, S., Lysenko, I., Senior, R., Börger, L., Bennett, D., Choimes, A., Collen, B., Day, J., De Palma, A., Diaz, S., Echeverria-Londono, S., Edgar, M., Feldman, A., Garon, M., Harrison, M., Alhusseini, T., Purvis, A., 2015. Global effects of land use on local terrestrial biodiversity. Nature 520, 45–50. doi:10.1038/nature14324.

Orynbaikyzy, A., Gessner, U., Conrad, C., 2019. Crop type classification using a combination of optical and radar remote sensing data: a review. Int. J. Remote Sens. 40, 6553–6595. URL: https://doi.org/10.1080/01431161.2019.1569791, doi:10.1080/01431161.2019.1569791, arXiv:https://doi.org/10.1080/01431161.2019.1569791.

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.

Pereira, R.C., Santos, M., Rodrigues, P., Henriques Abreu, P., 2020. Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes. J. Artif. Intell. Res. 69, 1255–1285. doi:10.1613/jair.1.12312.

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.

Prendes, J., Chabert, M., Pascal, F., Giros, A., Tourneret, J.Y., 2015a. Change detection for optical and radar images using a Bayesian nonparametric model coupled with a Markov

random field, in: Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc., Brisbane, Australia. pp. 1513–1517.

Prendes, J., Chabert, M., Pascal, F., Giros, A., Tourneret, J.Y., 2015b. A new multivariate statistical model for change detection in images acquired by homogeneous and heterogeneous sensors. IEEE Trans. Image Process. 24, 799–812.

Prendes, J., Chabert, M., Pascal, F., Giros, A., Tourneret, J.Y., 2015c. Performance assessment of a recent change detection method for homogeneous and heterogeneous images. Revue Française de Photogrammétrie et de Télédétection 209, 23–29.

Quegan, S., Jiong Jiong Yu, 2001. Filtering of multichannel SAR images. IEEE Trans. Geosci. Remote Sens. 39, 2373–2379. doi:10.1109/36.964973.

Rabiner, L., 1989. A tutorial on Hidden Markov Models and selected applications on speech recognition. Proceedings of the IEEE 77, 257–286. doi:10.1109/5.18626.

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.

Rizzoli, P., Bräutigam, B., 2014. Radar backscatter modeling based on global tandem-x mission data. IEEE Transactions on Geoscience and Remote Sensing 52, 5974–5988. doi:10.1109/TGRS.2013.2294352.

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.

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. doi:https://doi.org/10.1016/j.compag.2021.106173.

Schölkopf, B., Tsuda, K., Vert, J.P., 2004. Kernel methods in computational biology. MIT Press, Cambridge, Mass.

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., Platt, J., 1999. Support vector method for novelty detection, in: Proc. NIPS, Denver, CO, USA. pp. 582–588.

Sedano, F., Kempeneers, P., Hurtt, G., 2015. A Kalman filter-based method to generate continuous time series of medium-resolution NDVI images. Remote Sens. 6, 12381–12408.

Segarra, J., Buchaillot, M.L., Araus, J.L., Kefauver, S.C., 2020. Remote sensing for precision agriculture: Sentinel-2 improved features and applications. Agronomy 10. doi:10.3390/agronomy10050641.

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.

Siachalou, S., Mallinis, G., Tsakiri-Strati, M., 2015. A hidden Markov models approach for crop classification: Linking crop phenology to time series of multi-sensor remote sensing data. Remote Sens. 7, 3633–3650. doi:10.3390/rs70403633.

Small, D., 2011. Flattening gamma: Radiometric terrain correction for SAR imagery. IEEE Trans. Geosci. Remote Sens. 49, 3081–3093. doi:10.1109/TGRS.2011.2120616.

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

Tirado, M., Clarke, R., Jaykus, L., McQuatters-Gollop, A., Frank, J., 2010. Climate change and food safety: A review. Food Res. Int 43, 1745–1765. doi:https://doi.org/10.1016/j.foodres.2010.07.003.

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.

Verbesselt, J., Hyndman, R., Zeileis, A., Culvenor, D., 2010. Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. Remote Sens. Environ. 114, 2970–2980. doi:https://doi.org/10.1016/j.rse.2010.08.003.

Verbesselt, J., Zeileis, A., Herold, M., 2012. Near real-time disturbance detection using satellite image time series. Remote Sens. Environ. 123, 98 – 108.

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.

Viovy, N., Saint, G., 1994. Hidden Markov models applied to vegetation dynamics analysis using satellite remote sensing. IEEE Trans. Geosci. Remote Sens. 32, 906–917. doi:10.1109/36.298019.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E.W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Contributors, S..., 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat. Methods doi:https://doi.org/10.1038/s41592-019-0686-2.

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.

Wang, K., Franklin, S., Guo, X., He, Y., Mcdermid, G., 2009. Problems in remote sensing of landscapes and habitats. Prog. Phys. Geog. 33, 747–768. doi:10.1177/0309133309350121.

Wegmuller, U., Cordey, R.A., Werner, C., Meadows, P.J., 2006. "Flashing Fields" in nearly simultaneous ENVISAT and ERS-2 C-band SAR images. IEEE Trans. Geosci. Remote Sens. 44, 801–805.

Wegmüller, U., Santoro, M., Mattia, F., Balenzano, A., Satalino, G., Marzahn, P., Fischer, G., Ludwig, R., Floury, N., 2011. Progress in the understanding of narrow directional microwave scattering of agricultural fields. Remote Sens. Environ. 115, 2423–2433.

Weiss, M., Jacob, F., Duveiller, G., 2020. Remote sensing for agricultural applications: A meta-review. Remote Sens. Environ. 236, 111402. doi:https://doi.org/10.1016/j.rse.2019.111402.

Wheeler, T., von Braun, J., 2013. Climate change impacts on global food security. Science 341, 508–513. doi:10.1126/science.1239402.

Whelen, T., Siqueira, P., 2018. Time-series classification of Sentinel-1 agricultural data over North Dakota. Remote Sens. Lett. 9, 411–420. doi:10.1080/2150704X.2018.1430393.

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.

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.

Zeng, L., Wardlow, B.D., Xiang, D., Hu, S., Li, D., 2020. A review of vegetation phenological metrics extraction using time-series, multispectral satellite data. Remote Sens. Environ. 237, 111511. doi:https://doi.org/10.1016/j.rse.2019.111511.

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.

Zhao, Y., Nasrullah, Z., Li, Z., 2019. PyOD: A Python toolbox for scalable outlier detection. J. Mach. Learn. Res. 20, 1–7. URL: http://jmlr.org/papers/v20/19-011.html.

Zhou, Z.G., Tang, P., 2016. Continuous anomaly detection in satellite image time series based on z-scores of season-trend model residuals, in: Proc. IEEE IGARSS, Beijing, China. pp. 3410–3413.

Zhou, Z.G., Tang, P., Zhou, M., 2016. Detecting anomaly regions in satellite image time series based on sesaonal autocorrelation analysis, in: Proc. ISPRS, Prague, Czech Republic. pp. 303–310. URL: https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/III-3/303/2016/, doi:10.5194/isprs-annals-III-3-303-2016.