Missing data reconstruction and anomaly detection in crop development using agronomic indicators derived from multispectral satellite images

Mohanad Albughdadi, Denis Kouamé, Guillaume Rieu, Jean-Yves Tourneret

MISSING DATA RECONSTRUCTION AND ANOMALY DETECTION IN CROP DEVELOPMENT USING AGRONOMIC INDICATORS DERIVED FROM MULTISPECTRAL SATELLITE IMAGES

Mohanad Albughdadi[1], *Denis Kouamé*[1], *Guillaume Rieu*[3], *Jean-Yves Tourneret*[2]

[1] University of Toulouse, IRIT/UPS, France
[2] University of Toulouse, IRIT/INP-ENSEEIHT/TéSA, France
[3] TerraNIS, France
{mohanad.albughdadi, denis.kouame, jean-yves.tourneret}@irit.fr, guillaume.rieu@terranis.fr

ABSTRACT

This paper studies a new three-step procedure for detecting anomalies in crop development using temporal indicators derived from multispectral satellite images. These anomalies may result from seeding problems, heterogeneity, deficiency and stress. The first step estimates different biophysical and statistical parameters associated with these parameters from the observed images. In a second step, missing data that arise from the existence of clouds or limited coverage in the satellite image are reconstructed. Finally, the mean shift algorithm is used as an unsupervised classifier to detect anomalies in these reconstructed data. The proposed procedure is evaluated using agronomic indicators estimated from SPOT 5 Take 5 satellite images from the Beauce area in France.

***Index Terms*—** Agronomic indicators, Spot 5 satellite images, KNN reconstruction, anomaly detection, mean shift

1. INTRODUCTION

Remote sensing images can be used to estimate different agronomic indicators that have well defined structures dependent on the growth state of a crop. Overland is a software developed by Airbus Defense and Space that uses multispectral images to extract biophysical parameters of crops such as the fraction of green vegetation cover (fCover), the fraction of absorbed photosynthetically active radiation (FAPAR), the chlorophyll content (CHL) and leaf area index (LAI) [1]. Based on these parameters, agronomic advice can be delivered to farmers helping them to monitor their crops. The free access to Sentinel-2 images, characterized by a spectral richness and a fine temporal and spatial resolution will foster the development of image processing applications, in particular those related to crop development.

This paper investigates a three-step generic procedure to detect anomalies in crop development using the temporal evolution of indicators. The rest of the paper is organized as follows. The problem formulation and the algorithms used in the proposed procedure are described in Section 2. Results obtained on real data extracted from SPOT 5 satellite images are discussed in Section 3. Finally, some conclusions are drawn in Section 4.

This work was supported by the French FUI SparkinData project.

2. PROBLEM FORMULATION

The proposed unsupervised anomaly detection procedure contains three main components summarized in Fig. 1. These components are based on feature extraction, missing data reconstruction and unsupervised anomaly detection. These components are detailed in the following subsections.

2.1. Feature Extraction

Features are extracted from SPOT 5 multispectral images using the Overland software. The image collection is composed of 18 images acquired between April and September 2015 as part of the Take 5 experimentation in the Beauce area in France. Overland is used to extract biophysical parameters in 2400 wheat parcels. The selected biophysical parameters are fCover, FAPAR, CHL and LAI. Statistical indicators of these parameters (such as mean, variance, median) are then computed from the pixel values of each parcel. Note that a cloud detection procedure is applied in a pre-processing step, i.e., pixels with a reflectance higher than a given threshold are considered as cloud pixels. Biophysical parameters associated with cloud pixels are not computed yielding numerous missing data.

2.2. Reconstruction Methods

Missing data is a common problem that arises in many real world datasets. Missing data are generally due to erroneous data acquisition, incorrect measurements, absence of response, ... [2]. In our application, there are two main reasons for missing data when building temporal agronomic indicators. The first one is the existence of clouds that cover some parcels. The second reason is that some parcels are not

Fig. 1: A three-step procedure for anomaly detection in crop development.

covered in the acquisition process during specific days. Different approaches have been proposed in the literature to deal with missing data. These approaches can be categorized in two main classes. On the one hand, there are complete-case analysis methods that ignore all observations and/or variables associated with missing data [2]. These methods suffer from a loss of possibly critical information in the data. On the other hand, there are data reconstruction approaches that replace missing data with reasonable values, e.g., obtained by interpolation [3]. In what follows, we will focus on three missing data reconstruction methods from the second type: the mean imputation, the least squares reconstruction and the kNN reconstruction methods. For the ease of exposure, $\boldsymbol{X}$ denotes the data matrix of size $N_{\text{var}} \times N_{\text{obs}}$, where N_{var} and N_{obs} are the numbers of variables and observations, respectively. The i-th row of $\boldsymbol{X}$ is denoted as $\boldsymbol{X}_i^{\text{row}}$ while its j-th column is denoted as $\boldsymbol{X}_j^{\text{col}}$.

(i) *Mean Imputation*
This method is considered as the simplest method of missing data reconstruction. It simply replaces each missing value by the mean of the observed values remaining for that variable. However, this approach can severely modify the distribution of the missing variable, which may lead to complications such as underestimating the standard deviation and distortion of the relationships between the different variables by pulling correlation estimates toward zero.

(ii) *Least Squares Reconstruction*
The least squares (LS) approach described in [2] assumes that if an observation $\boldsymbol{X}_j^{\text{col}}$ of size N_{var} has N_k known variables, then $\boldsymbol{Y}_j^{\text{col}}$ is an observation of size N_k that can be expressed as

$$\boldsymbol{Y}_j^{\text{col}} = \boldsymbol{S}\boldsymbol{X}_j^{\text{col}} \tag{1}$$

where $\boldsymbol{S}$ is a sampling matrix of size $N_k \times N_{\text{var}}$. This matrix is extracted from the identity matrix of size $N_{\text{var}} \times N_{\text{var}}$ by removing the rows corresponding to the missing variables. From (1), it is straightforward to recover the missing data from the observations as follows

$$\boldsymbol{X}_j^{\text{col}} = \boldsymbol{S}^{\text{t}}\boldsymbol{Y}_j^{\text{col}}. \tag{2}$$

To reconstruct missing variables in $\boldsymbol{X}_j^{\text{col}}$, we can also define a matrix $\boldsymbol{S}_c$ using the removed missing rows from the identity matrix as discussed before. The following estimator can then be defined

$$\hat{\boldsymbol{X}}_j^{\text{col}} = \boldsymbol{S}^{\text{t}}\boldsymbol{Y}_j^{\text{col}} + \boldsymbol{S}_c^{\text{t}}\boldsymbol{V} \tag{3}$$

where $\boldsymbol{V}$ is a vector composed of the $N_{\text{var}} - N_k$ missing variables to be estimated. It is obvious that $\hat{\boldsymbol{X}}_j^{\text{col}}$ corresponds to estimating the vector $\boldsymbol{V}$ and then replacing the missing variables in $\boldsymbol{X}_j^{\text{col}}$. An estimation of $\boldsymbol{V}$ can be obtained by minimizing the energy of the second-order derivative of $\hat{\boldsymbol{X}}_j^{\text{col}}$. Therefore, $\boldsymbol{V}$ can be obtained by minimizing $\|\boldsymbol{D}\hat{\boldsymbol{X}}_j^{\text{col}}\|_2^2$, where $\boldsymbol{D}$ is the second order difference matrix of $\hat{\boldsymbol{X}}_j^{\text{col}}$, i.e., by solving the following problem

$$\min_{\boldsymbol{V}} \|\boldsymbol{D}(\boldsymbol{S}^{\text{t}}\boldsymbol{Y}_j^{\text{col}} + \boldsymbol{S}_c^{\text{t}}\boldsymbol{V})\|_2^2. \tag{4}$$

The solution of (4) is known to be

$$\widehat{\boldsymbol{V}} = -\left(\boldsymbol{S}_c\boldsymbol{D}^{\text{t}}\boldsymbol{D}\boldsymbol{S}_c^{\text{t}}\right)^{-1}\boldsymbol{S}_c\boldsymbol{D}^{\text{t}}\boldsymbol{D}\boldsymbol{S}^{\text{t}}\boldsymbol{Y}_j^{\text{col}}. \tag{5}$$

(iii) *kNN Reconstruction*
The kNN reconstruction defined in [4] is an extension of the kNN algorithm, which finds the k most relevant complete observations using the Euclidean distance and weights the contribution of each observation in the missing data. Assuming that the two observations $\boldsymbol{X}_i^{\text{col}}$ and $\boldsymbol{X}_j^{\text{col}}$ have the same size, the distance between them can be calculated as follows

$$d\left(\boldsymbol{X}_i^{\text{col}}, \boldsymbol{X}_j^{\text{col}}\right) = \left(\frac{\sum_{l=1}^{N_{\text{var}}} r_l^i r_l^j \sqrt{(x_l^i - x_l^j)^2}}{\sum_{l=1}^{N_{\text{var}}} r_l^i r_l^j}\right)^{1/2} \tag{6}$$

where $\boldsymbol{X}_i^{\text{col}} = \{x_l^i, l = 1, \ldots, N_{\text{var}}\}$, $\boldsymbol{X}_j^{\text{col}} = \{x_l^j, l = 1, \ldots, N_{\text{var}}\}$ and where the binary variable r_l is defined such that $r_l = 0$ if $x_l = \text{NaN}$ and $r_l = 1$ if $x_l \neq \text{NaN}$. Eq. (6) replaces the ℓ_2 distance used in the kNN classification algorithm and accounts for the missing values. After finding the k nearest neighbors and sorting the distances calculated using (6), the missing values can be reconstructed as follows

$$\widehat{x_l^j} = \sum_{v=i}^{k} \omega_v^j x_l^v \tag{7}$$

where the weight ω_v^j is defined as

$$\omega_v^j = \frac{1/d(\boldsymbol{X}_j^{\text{col}}, \boldsymbol{X}_v^{\text{col}})}{\sum_{v=i}^{k}\left(1/d(\boldsymbol{X}_j^{\text{col}}, \boldsymbol{X}_v^{\text{col}})\right)}. \tag{8}$$

Note that the corresponding algorithm for kNN reconstruction is summarized in Algorithm 1.

2.3. Unsupervised Classification

At this stage, the reconstructed temporal indicators are used to detect the abnormal agricultural parcels at a specific time instant. To achieve this objective, we rely on the mean shift algorithm[1]. The mean shift algorithm is a robust feature space

[1]Note that other clustering algorithms could be used instead of the mean shift algorithm.

Algorithm 1 kNN Reconstruction Algorithm

Input: Incomplete dataset $\boldsymbol{X}$, $k-$nearest neighbors
Output: Imputed dataset $\boldsymbol{X}'$
for each observation $\boldsymbol{X}_j^{\text{col}} \in \boldsymbol{X}$ **do**
  for each missing value $x_l^j, l = 1, \ldots, N_{\text{var}}$ **do**
    Find the $k-$nearest neighbors using (6).
    for each $x^v, v = 1, \ldots, k$ close to x^j **do**
      Calculate the weight ω_v^j according to (8).
    Estimate x_l^j according to (7).
  Return $\boldsymbol{X}'$.

algorithm that has been widely used for clustering and classification [5]. It is a non-parametric iterative algorithm that relies on a kernel density estimation and does not require to set the number of classes. This number is automatically estimated by estimating the number of modes of a multivariate distribution underlying the feature space [5]. Dense regions in the feature space correspond to modes of the underlying probability density function (pdf). The mean shift algorithm assigns each data point to the closest peak of the pdf by defining a kernel around each data point and computing its mean. The center of the kernel is then shifted to the mean in an iterative procedure until convergence. Denoting as N_{obs} the number of data points $\boldsymbol{X}_j^{\text{col}} \in \mathbb{R}^{N_{\text{var}}}$, $j = 1, ..., N_{\text{obs}}$ and assuming that each of these data points is associated with a bandwidth $h_j > 0$, the mean shift vector is defined as

$$\boldsymbol{m}_G(\boldsymbol{X}^{\text{col}}) = \frac{\sum_{j=1}^{N_{\text{obs}}} \frac{1}{h_j} \boldsymbol{X}_j^{\text{col}} g\left(\left\| \frac{\boldsymbol{X}^{\text{col}} - \boldsymbol{X}_j^{\text{col}}}{h_j} \right\|^2\right)}{\sum_{j=1}^{N_{\text{obs}}} \frac{1}{h_j} g\left(\left\| \frac{\boldsymbol{X}^{\text{col}} - \boldsymbol{X}_j^{\text{col}}}{h_j} \right\|^2\right)} - \boldsymbol{X}^{\text{col}} \quad (9)$$

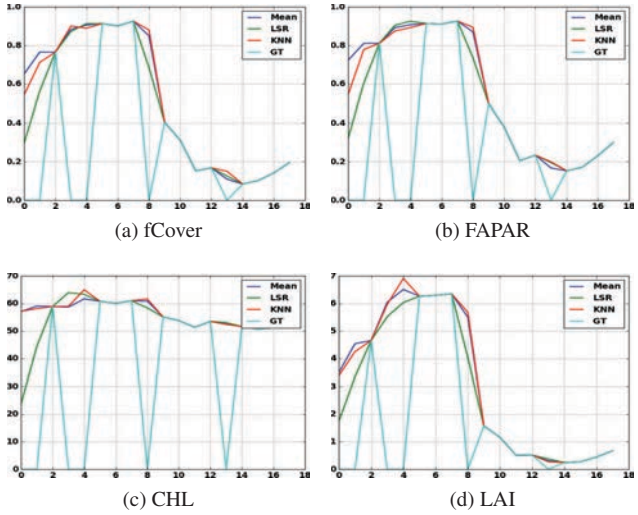
where $g(.) = -k'(.)$ and k' is the derivative of the kernel profile k. In (9), the bandwidth is estimated using the nearest neighbors of $\boldsymbol{X}^{\text{col}}$ [6]. Denoting as $\boldsymbol{X}_{j,k}^{\text{col}}$ the k nearest neighbor of $\boldsymbol{X}_j^{\text{col}}$, the bandwidth h_j can be computed using the ℓ_1 norm as

$$h_j = \|\boldsymbol{X}_j^{\text{col}} - \boldsymbol{X}_{j,k}^{\text{col}}\|_1. \quad (10)$$

Finally, data points that converge to similar values are considered to be in the same class.

3. SIMULATION RESULTS

The proposed anomaly detection procedure has been validated using 18 SPOT 5 Take 5 satellite images. These images were acquired in the period between 10-th April to 17-th September 2015 and consist of 2400 unique parcels. Different biophysical parameters were extracted from these images and some statistical indicators were then computed. In the following simulations, these statistical indicators are defined as the means of the fCover, FAPAR, CHL and LAI parameters, as an example. For each indicator, a dataset was constructed by extracting the values associated with each parcel from the 18 days. If the parcel was not acquired at a specific date, the indicator value was considered as missing. The resulting data matrix for each indicator is of size $N_{\text{var}} \times N_{\text{obs}} = 18 \times 2400$. Before applying the unsupervised anomaly detection procedure, the three aforementioned methods of missing data reconstruction were compared. For the kNN reconstruction algorithm, the number of nearest neighbors (kNN) was set to $k = 50$ using cross validation. Note that choosing a value of this parameter between 10 and 100 gives similar results. Fig. 2 shows different examples of reconstructed data using these methods. To further analyze



(a) fCover (b) FAPAR

(c) CHL (d) LAI

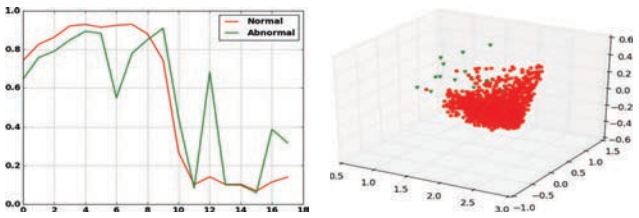
Fig. 2: Performance of the missing data reconstruction methods for four different agronomic indicators. Mean, LSR, KNN and GT refer to mean imputation, kNN reconstruction, least squares reconstruction and ground truth, respectively.

these results, the reconstructed datasets obtained with the four indicators using the kNN reconstruction algorithm were considered as a ground truth. In a second step, 50% of each observation were randomly set to zero in order to mimic a realistic scenario of missing data. Each of the reconstruction methods was run 100 times to reconstruct the missing data in the aforementioned indicators and the averaged root mean square errors (RMSEs) were estimated. These values are reported in Table 1. From these results, it is clear that the kNN reconstruction algorithm gives a better performance when compared to the other two methods. To evaluate the performance of the unsupervised anomaly detection method, the fCover dataset after kNN reconstruction was clustered using the mean shift algorithm described in Section 2.3. To estimate the bandwidth, the number of nearest neighbors was set to $k = 50$ using cross validation. Due to the sparsity of the data

Table 1: Estimated average RMSEs for 100 runs between the ground truth datasets and the three reconstruction methods for 4 different indicators.

Indicator	Mean	LSR	kNN
fCover	0.062	0.149	**0.056**
FAPAR	0.052	0.153	**0.047**
CHL	1.577	8.847	**1.474**
LAI	0.595	1.000	**0.558**

in high-dimensional space, this parameter should be large enough to ensure that all data points have neighbors within a distance h. The mean shift algorithm manages to identify two different classes whose centers are shown in Fig. 3[a]. A 3-D projection of the dataset using 3 principal components (obtained by a principal component analysis (PCA)) is shown in Fig. 3[b]. The red dots and green triangles in this figure correspond to normal and abnormal parcels, respectively. Fig. 4 shows the fCover temporal indicator of one of the abnormal parcels whereas the corresponding temporal evolution of the fCover pixel values in 14 days (the other 4 days are missing data) is shown in Fig. 5. A visual comparison between the fCover indicator of this parcel and the center of the normal class (Fig. 4) reveals the abnormal behavior of the parcel. Indeed, this parcel shows a state of crop senescence in an early stage of crop development (in days 1 to 3 of Fig. 5), which may affect the crop yields. Moreover, a strong heterogeneity appears in this parcel (in days 4 to 8 of Fig. 5), which is probably caused by a poor agronomic practice such as sowing density.



(a) Class centers (b) 3-D projection

Fig. 3: Normal and abnormal class centers [a]; 3-D projection of the fCover temporal indicators associated with the parcels [b].

4. CONCLUSION

This paper studied a procedure for detecting abnormal agricultural parcels from satellite images. This procedure consisted of three main steps devoted to feature extraction, missing data reconstruction and unsupervised anomaly detection. The results obtained with the proposed procedure on agronomic indicators are promising. In the reconstruction step, the kNN reconstruction algorithm leads to very good performance when compared to the mean imputation and least squares reconstruction. The kNN reconstruction algorithm was used to reconstruct missing temporal indicators. However, it could also be used to reconstruct multiple indicators acquired during the same day since it relies on known values captured from neighboring observations. The last step was the mean shift algorithm allowing normal and abnormal parcels to be classified. Future work will focus on applying this framework to Sentinel-2 data. Another interesting prospect is the application of this framework to a more concrete agricultural application such as irrigation monitoring and yield prediction.

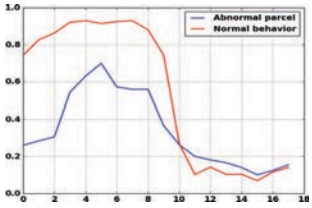


Fig. 4: Mean fCover indicator of an abnormal parcel after the reconstruction procedure along with the center of the normal parcels.

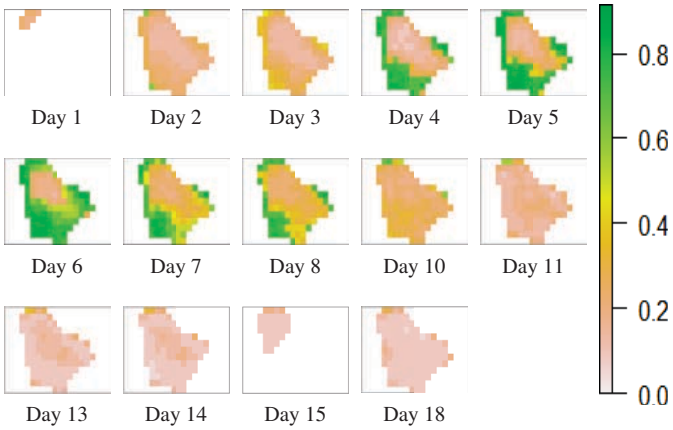


Fig. 5: Temporal evolution of the fCover indicator for an abnormal parcel.

5. REFERENCES

[1] H. Poilvé, "Towards an operational GMES land monitoring core service - BIOPAR methods compendium - MERIS FR biophysical; Geoland2," Tech. Rep., Immenstaad, Germany, February 2010.

[2] R. Little and D. Rubin, *Statistical analysis with missing data*, John Wiley & Sons, 2014.

[3] G. Batista and M. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 519–533, 2003.

[4] O. Troyanskaya et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[5] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, 2002.

[6] B. Georgescu et al., "Mean shift based clustering in high dimensions: A texture classification example," in *Proc. ICCV*, Nice, France, 2003, IEEE, pp. 456–463.