Field heterogeneity detection based on the modified FastICA RGB-image processing

L Mirvakhabova\(^1\), M Pukalchik\(^1\), S Matveev\(^1,2\), P Tregubova\(^1\) and I Oseledets\(^1,2\)

\(^1\) Skoltech Center for Computational and Data-Intensive Science and Engineering, Skolkovo Institute of Science and Technology, 1 Nobel Street, Moscow, 121205, Russia
\(^2\) Marchuk Institute of Numerical Mathematics, RAS, 8 Gubkin Street, 119333, Moscow, Russia

M.Pukalchik@skoltech.ru

Abstract. In the light of the increasing demand for food production, climate change challenges, and economic pressure, precision farming is becoming an ever-growing market. Nowadays using satellite images for field zoning has become a widely spread technique employed in precision agriculture. Usually satellite images are used to evaluate Normalized Difference Vegetation Index (NDVI) as a common yield assessment. At the same time, it is rather a challenging task to design NDVI maps as compared to ordinary RGB images. We propose a new algorithm for site-specific zoning based on the performance of a well-known and widely used ICA algorithm (FastICA). We analyse high-resolution RGB satellite images for 3 arable fields located in Kursk region, Russia, (2017) provided by the Planet Labs service. The main outcomes from this study are (i) the algorithm creates site-specific zoning maps with a relative accuracy of yield distribution maps between 0.78 and 0.89; (ii) the algorithm requires a relatively small dataset – from 8 to 10 RGB images. The obtained results indicate that the algorithm in question can be used not only to identify management zones but also to map the variations in yield within the fields.

1. Introduction
Digital technologies and Big Data have recently become an intrinsic part of our life and come into use in practically all spheres with no exception of agriculture. Among other techniques, it is the site-specific field zoning that has become a crucial step in precision agriculture management [1] and came into a considerable use. Nowadays that satellite data have become more accessible [2] owing to a large number of sources for the high-resolution imagery such as Planet, RapidEye, and Sentinel-2 satellites, many studies started to portray the benefits obtained from the use of high-resolution satellite imagery for identifying within-field yield variation [3–5] and mapping management zones [6]. Meanwhile, at least two issues hamper the spreading of automatic site-specific field delineation techniques to be used by farmers.

Firstly, the Normalized Difference Vegetation Index (NDVI) is one of the most widely used approaches to assess yield variation within the field is to use [7]. Though the assessment of crop production with the help of NDVI has several restrictions and limits, this index has become the grounds for elaborating many vegetation indexes for crop monitoring [8]. In particular, calculations of NDVI require to obtain the data from the sensor operating in the near infrared (NIR) wavelength.
At the same time, though with the advent of reliable and low-cost drones the ability to create vegetation maps for farm fields had opened up for everyone. But the most standard low-cost drones have only the cameras which can capture red, green, and blue (RGB) bands of light to produce images, and it may be too expansive to upgrade these to NIR spectra. Meanwhile, only a few studies have focused on the advantages provided by utilizing high-resolution RGB imagery to serve this purposes.

Secondly, several well-known commercial products like Cropio, ClearAg and AgPixel can provide the services for farmers. Unfortunately, the companies are usually reluctant to fully disclose the detailed information about the algorithms used in their analysis to their customers. Another thing is a very limited number of farmers, especially those living in countries with a low income, can afford such products.

Turning back to Independent Component Analysis (ICA), it attracted attention owing to a wide range of its potential applications [9]. The goal of ICA is to recover independent sources which were given only as pictures of unknown linear mixtures of unobserved independent source signals. In contrast to correlation-based transformations, such as Principal Component Analysis (PCA), ICA not only decorrelates signals but also reduces higher-order statistical dependencies in the attempt to make the signals as independent as possible. Recently ICA was proposed as a tool for sorting out unmixed hyperspectral data [10, 11].

Since ICA algorithm includes higher statistics (as compared to the second order statistics), it seems to be a perfect way to reduce dimensionality. To be able to obtain the generating factors, ICA was designed in such a way not to search for the principal components, which represents the maximum of return dispersion, but for much more independent factors, which enables it to linearly generate the returns. ICA has proven to show high performance as compared to PCA in various fields, such as object recognition and geoscience applications. However, the cases of using ICA in spectral data analysis in dimensionality reducing source separation and in data compression are regrettably rare.

This study explores the methodology and the algorithm for approximating RGB satellite images to create site-specific zone maps basing on the FastICA.

2. Material and Methods
To for the sake of convenience the paper is divided into several major steps (figure 1); namely: 1) dataset collection; 2) image alignment using standard techniques for border detection; 3) running and adapting FastICA algorithm for site-specific map creation; 4) post-processing step, to compare the obtained results with a real yield map.

The collected dataset (images) was analyzed with the help of the authored algorithm based on the FastICA. In this study we use the FastICA algorithm of Hyvärinen [12] and Hyvärinen et al. [13]. The FastICA algorithm consists of an initial pre-whitening stage and an iteration stage. Application of methods for high-resolution satellite image processing has also been studied by Kozoderov and Dmirtiev [14] in a series of works, the main target area of which was the consideration of heterogeneity in forests for differentiation between trees species. In our work the target application area is heterogeneity detection for agro-fields and a possible detection of hazardous zones in soils. The goal is to find the components that are maximally independent and non-Gaussian [15, 16]. The main advantage of the proposed algorithm is it can measure a system or a phenomenon without designing any experimental conditions. Moreover, FastICA can be used to investigate the structure of data in cases when working hypotheses are not proved or they are considered to be too constrained or simplistic [12]. (For further details concerning algorithm see ‘Results’ and ‘Discussion’).

Finally, we compare the received zone-specific maps with in-situ data on crop yields obtained from the farmers. We use the coefficient of determination $R^2$ [17] and Frobenius norm as quality measure of the prediction; its positive value evidence that there is high probability to fulfill an accurate prediction.

All of the mentioned algorithms and their modifications are implemented using Python libraries, namely, standard pure fast-ICA and PCA as the method of scikit.learn library.
Figure 1. Flow chart of the process.

Validation dataset

The three fields located in a forest steppe zone of Russia (Kursk oblast) and cultivating different types of crops were tested in the course of the research. The region is characterised by an undulating plain relief with mild-slope gullies and unshaped depressions. The climate is moderately cold, with mean annual precipitation of 350–570 mm and mean annual air temperature of 5.4 °C. Average annual precipitation is about 545 mm. Average air temperature is -7.2 °C in January and +20.1 °C in June. Vegetation period lasts 188 days. During winter months’ soil freezes down to the depth of 69 cm.

Table 1. Descriptive information about validation fields (crop type, total area) and satellite data (imagery acquisition date, the image size in pixels).

| Field | Images | Area, ha | Crop type, 2017 | Image size | Number of tested images per each month |
|-------|--------|----------|----------------|------------|---------------------------------------|
|       |        |          |                | April      | May | June | July | August | September |
| 1     | ![Image](field1.png) | 90        | Soybean        | 291×451    | 3   | 2    | 4    | 2      | 11        | 11        |
| 2     | ![Image](field2.png) | 210       | Spring wheat   | 367×621    | 5   | 2    | 4    | 3      | 6         | 13        |
| 3     | ![Image](field3.png) | 126       | Winter wheat   | 334×599    | 2   | 2    | 4    | 1      | 7         | 2         |
In the course of the research we operated a set of satellite images (RGB) retrospectively collected from Planet Labs service [18]. Manual selection of images was performed to avoid cloud-covered scenes and to determine specific dates which could represent the emerging and the ripening of crops. For more detailed information about the fields and the corresponding sets of images see table 1.

3. Results and Analysis

3.1. Image alignment with standard techniques for border’s detection
For each field we discovered several independent components from the set of aligned satellite images. The number of these components varied from 1 to N, where N is the size of the dataset. At that we fixed the number of independent components as a vital prerequisite for a subsequent correct comparison. Then we made all calculations applying FastICA algorithm to all possible sizes of the dataset, and thus we obtained the reconstruction error (i.e. the accuracy in relative Frobenius norm) between our zone-specific map and real yield maps. Finally, we decided to choose the smallest value for N, which corresponded to a reasonable precision of input data approximation. In case with our dataset this value was equal to 6.

It is worth mention here that applying the modified Fast-ICA directly to satellite images from dataset collection failed to be a success, since these images required a border alignment of methods than the standard borders detection methods that we used for to solve this problem [19]. In figure 2A there is the processed image without the border alignment; as we can see, the only detected heterogeneity is concentrated as noise mainly along the border of the field. Since a straightforward application of ICA appeared to lead to such unsatisfactory results, a strong need in image alignment existed. Figure 2B presents the results of processing the aligned dataset.

![Image Alignment](image.png)

Figure 2. The results of images preprocessing using the modified FastICA algorithm: [A] without border alignment procedure, [B] with border alignment procedure.

3.2 Rank-1 correction and analysing ICA-coefficients
Let us denote the normalized matrix as $\tilde{A}$, so for each row we have

$$\tilde{A} = \frac{A_i}{\|A_i\|}$$

In a matrix form:

$$\tilde{A} = \Lambda A,$$

$$\Lambda \in \mathbb{R}^{n \times n},$$

$$\Lambda = \text{diag} \left( \frac{1}{\sum_{i=1}^{m} a_{ii}}, \frac{1}{\sum_{i=1}^{m} a_{ij}}, \ldots, \frac{1}{\sum_{i=1}^{m} a_{ik}} \right).$$
The coefficient matrix reflects the time dynamics, but if we study the properties of the obtained coefficients immediately, we may face the fact that the coefficients may look the same – up to a scale. Therefore, we need to reduce the influence of common patterns in each coefficient time series. In order to separate the intrinsic coefficient patterns, we subtract a rank-1 approximation:

\[ \bar{A} = \mathcal{U} \sum V^T, \]

\[ A' = \bar{A} - \sigma_1 u_1 \otimes v_1^T, \]

where \( \otimes \) stands for Kronecker product operation. Let us denote the obtained matrix of coefficients as \( \bar{A} \) with the rows

\[ \bar{a}_i = \frac{1}{\|a_i\|} a_i - \sigma_1 u_1 v_1^T, \]

where \( a_i \) is the row of \( A \).

It appeared that if we chose a wrong number of components (more than we needed), part of components vectors would have almost the same values. Let us suppose, for example, that \( q \) of \( k \) are the components of the same value and for the sake simplicity denote them as \( s_1 \). Then the element of \( X \) matrix runs as follows:

\[ X_{ij} = \bar{a}_i^T s_j = \left( \frac{1}{\|a_i\|} a_i - \sigma_1 u_1 v_1^T \right)^T s_j. \]

and the row is as follows:

\[ X_i = \bar{a}_i^T (s_1, ..., s_q, s_{q+1}, ..., s_k). \]

Summing up, this procedure is applied to extract both all the independent components and the corresponding coefficient matrix with a reduced influence on the common trend. Since we have performed a normalization step, a natural measure of heterogeneity is the standard deviation. We suggest that the components with higher standard deviation would be more informative.

The results of these procedure in case of field 2 is shown in figure 3. Fast-ICA coefficients (independent component 1–6) without rank-1 correction appeared to have common time-dynamics patterns (figure 3A); while figure 3B presents coefficients for applying rank-1 correction: it seems that the peak located between 03/08 and 21/08 may be tractable. In fact, however, this period of time was related to spring wheat harvesting dates.

![Figure 3](image)

**Figure 3.** Fast-ICA coefficients before and after preprocessing field 2 (as an example). The lines of different kind (1–6) correspond to different independent components calculated for each satellite images (dates) to be analysed by modified FastICA algorithm.
3.3. Final discussion of method performance
To evaluate the performance of the proposed method we compared the obtained components with in-situ collected yield data on the selected sites. The first step was to normalize values for pixels of ICA-components and to yield maps into [0,1] interval.

The following measurements: a determination coefficient (figure 3) and a relative Frobenius norm (table 2) were used to evaluate the prediction accuracy. According to the results of the calculation, the determination coefficient ($R^2$) appeared to be low in Field 1 of our dataset; thus we included accuracy measurements in Frobenius norm. To find the value of a relative Frobenius norm we first calculated the difference between two normalized image matrices; to be more particular the first one corresponded to the chosen component, while the second one was a yield map. After that we calculated the ratio of Frobenius norm for the difference matrix to the norm for the yield map matrix. As can be seen, the accuracy of the modified FastICA appeared to be higher than that of PCA (which is the method considered to be a natural and wide-spread alternative to the one proposed in the paper).

Table 2. The comparison between the predictions of applying two methods to the validation dataset.

| Field | Relative Frobenius norm |
|-------|--------------------------|
|       | Modified FastICA | PCA |
| 1     | 0.79                   | 0.65 |
| 2     | 0.89                   | 0.86 |
| 3     | 0.78                   | 0.73 |

Figure 4. Comparison between the yield map and the calculated maps using modified FastICA.
4. Conclusion
The present work evaluates the ways of applying Independent Component Analysis (ICA) to crop yield estimation and feature extraction.

This research can be applied in several ways. Thus, the obtained results revealed that the independent components discovered in the course of research corresponded to yield maps (with accuracy scores in Frobenius norm between 0.78 and 0.89 and R² values between 0.40 and 0.92). In addition, the proposed method can be employed in industry to detect unprofitable areas which seems to be more affordable than using additional equipment. Moreover, the elaborated method may allow us to predict yield variability during the cropping period. This is a great advantage of the method as compared to using yield maps which can be collected only at harvest.

The main drawback of this study is a small-sized dataset. Hence, the outlook for the further research is to proceed with collecting the data: first and foremost, to increase the number of yield maps for verification.

Acknowledgements
The research was supported by Skoltech Next Generation Program (NGP).

References
[1] Oliver MA (ed) Geostatistical 2010 Applications for Precision Agriculture. Springer Netherlands
[2] Azzari G, Jain M and Lobell D B 2017 Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries Remote Sensing of Environment 202 129–41
[3] Yang C, Everitt J H, Du Q, Luo B, Chanussot J 2013 Using high-resolution airborne and satellite imagery to assess crop growth and yield variability for precision agriculture Proceedings of the IEEE 101 no3 582–592
[4] Schut A G T, Traore P C S, Blaes X and de By R A 2018 Assessing yield and fertilizer response in heterogeneous smallholder fields with UAVs and satellites Field Crops Research 221 98–107
[5] Seifert C A, Azzari G and Lobell D B 2018 Satellite detection of cover crops and their effects on crop yield in the Midwestern United States Environmental Research Letters 13 064033
[6] Novoa J, Chokmani K and Lhissou R 2018 A novel index for assessment of riparian strip efficiency in agricultural landscapes using high spatial resolution satellite imagery Science of The Total Environment 644 1439–51
[7] Rouse J W, Haas R H, Schell J A and Deering D W 1974 Monitoring vegetation systems in the Great Plains with ERTS, In S C Freden, E P Mercanti and M Becker (eds) Third Earth Resources Technology Satellite–1 Syposium Volume I: Technical Presentations NASA Washington, D.C. 309–317
[8] Xue J and Su B 2017 Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications Journal of Sensors 2017 1–17
[9] Comon P 1994 Independent component analysis – a new concept? Signal Processing 36(3) 287–314
[10] Parra LC et al 2000 Unmixing Hyperspectral Data Proc. Adv. Neural Inf. Process. Syst. 12 942–948
[11] Chen Y G and Qian S E 2011 Denoising of Hyperspectral Imagery Using Principal Component Analysis and Wavelet Shrinkage IEEE Transactions on Geoscience and Remote Sensing 49(3) 973–80
[12] Hyvärinen A 1999 Fast and robust fixed-point algorithms for independent component analysis IEEE Transactions on Neural Networks 10(3) 626–34
[13] Hyvärinen A 2013 Independent Component Analysis: Recent Advances. Philosophical Transactions of the Royal Society A 371 20110534
[14] Kozoderov V V and Dmitriev E V 2008 Remote sensing of soils and vegetation: regional aspects
    *International Journal of Remote Sensing* **29** (9) 2733–48
[15] Hyvärinen A, Ramkumar P, Parkkonen L and Hari R 2010 Independent component analysis of
    short-time Fourier transforms for spontaneous EEG/MEG analysis *NeuroImage* **49**(1) 257–71
[16] Shimizu S 2012 Joint estimation of linear non-Gaussian acyclic models *Neurocomputing* **81** 104
[17] Anderson-Sprecher R 1994 Model Comparisons and $R^2$ *The American Statistician* **48** 113–17
[18] Planet Team 2017 Planet Application Program Interface: In Space for Life on Earth (San
    Francisco, CA)
[19] Lowe D G 1999 Object recognition from local scale-invariant features, *Proceedings of the
    International Conference on Computer Vision* **2** 1150–57