Applying a Principle Component Analysis to Search for Objects on Historical Territories by the Spectral Brightness of Vegetation

A G Zlobina and I V Zhurbin
Udmurt Federal Research Center, Ural Branch of the Russian Academy of Sciences,
34 Baramzina str., Izhevsk, 426067, Russia
E-mail: elf54@yandex.ru

Abstract. An approach is proposed for analyzing multispectral aerial photography data to identify traces of human activity; this approach is based on the application of a principal component analysis. Its efficiency is illustrated by a case study of historical territory – the Pudemsky Ironworks (Udmurt Republic, Russia). About 67% of the arable land area abandoned in the latter half of the 19th century is revealed. Nowadays, this site is covered with forest vegetation and no longer has any striking visual or spectral differences from the environment.

1. Introduction
The late 18th – early 19th centuries was a period of rapid development of the mining industry of the Urals and Cis-Urals. Mining districts were formed, which are historically developed elaborate territorial socioeconomic networks with a multistructure system of factory economy. Apart from industrial facilities, they also included huge land with raw materials, fuel and energy resources, villages of ascribed peasants, arable land, meadows, etc. The use of natural resources was extensive, which ultimately led to rapid mineral depletion and deforestation. This is one of the main causes of the crisis in this industry in the following period.

To date, the territories belonging to the mining districts have lost their external distinctive features. Therefore, traces of human activity can only be revealed by indirect signs — a change in vegetation within the local site. The reason for this is the difference in the amount of moisture and nutrients that plants extract from the soil. For instance, stunted vegetation grows over the foundations of extinct structures; whereas dense and lush vegetation, over ancient pits or ditches filled with humus [1]. Such sites are distinguished not only by vegetation cover condition but by spectral brightness as well. Consequently, traces of anthropogenic impact can be revealed using multispectral aerial photography data. As a rule, the identification of such areas is carried out via visual analysis of images in different spectral channels [2], by designing distribution maps of vegetation indices [3, 4] and using methods of automatic image analysis (in most cases, cluster analysis) [5–7].

The purpose of this study is to develop and evaluate a method of searching for areas of anthropogenic impact in the 18th–19th centuries (arable land, cutover areas, mines, etc.). In the latter half of the 19th century, once these sites were no longer in use, the process of secondary succession began. At present, the initial state of the vegetation cover has been practically restored. Thus, these
areas do not have any striking visual and spectral differences from the natural environment. Consequently, the traditional approach for analyzing multispectral aerial photography data is ineffective in revealing areas of anthropogenic impact in historical territories.

The authors’ major contribution is that they’ve developed a fundamentally different way to solve this problem. By and large, this approach is well-known and based on the application of principal component analysis (PCA). It is known that the visualization of the first principal components that explain most of dispersion of the source data offers a generalized representation of the site in all spectral channels. All natural objects that are steadily visible in several source images stand out on these components. From this perspective, PCA makes it possible to avoid the redundancy of the source data, by replacing the analysis of the entire set of images with the analysis of the first principal components. The last principal components contain information about hidden patterns — local features of the site that are not overtly registered in the source data. These features do not have a marked contrast; therefore, they can be interpreted as traces of anthropogenic impact — the ones that have been significantly “smoothed out” by secondary succession. That is, as far as the problem being solved is concerned, which is to identify minor changes in the vegetation cover, an analysis of the last principal components will be of great interest for research.

2. Processing method

The source images obtained via multispectral aerial photography are digital images. The data is correlated, which means that with an increase in the pixel brightness value in one spectral channel, the brightness values in others increase as well. Using the PCA method makes it possible to eliminate correlation dependencies.

Let us introduce the following notation: \( m \) is the number of features (spectral channels); \( n \) is the number of pixels (image); \( \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T \) is a column vector comprised of the pixel brightness values in the \( i \) spectral channel; \( \mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) \) is a matrix of the source data; \( \mathbf{Y} = (y_1, y_2, \ldots, y_m) \) is a matrix of principal components.

The algorithm for converting the source data into principal components is carried out according to the following scheme [8].

Step 1. Standardization of evidence:

\[
\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T \Rightarrow \mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{im})^T, \text{ where } z_{ij} = (x_{ij} - \bar{x}_i)^2 / s_j, \text{ for } i = 1, \ldots, m, j = 1, \ldots, n.
\]

Step 2. Calculation of eigenvectors \( \Lambda = \{\alpha_i\}_{i=1}^m : \alpha_i = (\alpha_{i1}, \ldots, \alpha_{im})^T \) and eigenvalues \( \{\lambda_1, \lambda_2, \ldots, \lambda_m\} \) of the correlation matrix of standardized source data. Ordering eigenvalues in descending order:

\[
\lambda_1 \geq \lambda_2 \ldots \geq \lambda_m.
\]

Step 3. Determination of the matrix of principal components: \( \mathbf{Y} = \mathbf{Z} \Lambda \).

The principal components obtained as a result of this conversion are orthogonal (pairwise linearly independent) and ordered by decreasing dispersion. Further analysis of the images will be reduced to principal component analysis.

3. Field experiment, results and discussion

Research was carried out on the historical territory of the Pudemsky ironworks (18th – 19th centuries, Pudem village, Udmurt Republic). According to a map of Glazovsky Uyezd, the Vyatka Governorate, made in 1825, the site of multispectral aerial photography (figure 1a, red outline) was a forest, with central part cleared for plowing (figure 1a, orange hatching). After the plant had been closed, the abandoned arable land was again overgrown with woodland. At present, it has no external distinguishing features either during an on-site investigation or in photographs in the visible range (figure 1b). There is only the road crossing the site from west to east that has remained.
Figure 1. Pudemsky ironworks: a) a map of Glazovsky Uyezd, the Vyatka Governorate, made in 1825; b) an image in the visible range.

According to the results of multispectral aerial photography (figure 2), the images of the studied surface were obtained in green (G, 530–570 nm), red (R, 640–680 nm), red-edge (RE, 730–740 nm), and near-infrared (NIR, 770–810 nm) spectral channels (Finko LLC, Izhevsk). The imaging was made in May 2018 – during the early growing season. Using a priori information about the spectral curves of forest vegetation (figure 3a), it can be established that the site is a mixed forest, with central part being mostly coniferous. Same is the interpretation of the distribution maps of normalized difference vegetation index (NDVI) and ratio vegetation index (RVI) [9] that allow separating the vegetation cover from the open soil in a more contrast way (figure 3b).

Thus, the traditional analysis of multispectral aerial photography data did not reveal any traces of human activity. Only the type of forest vegetation was determined. Therefore, to uncover hidden dependencies, the original images were converted to principal components. The visualization of principal components is shown in figure 4. The corresponding eigenvectors and eigenvalues of the correlation matrix are listed in tables 1 and 2.

The first principal component of PC1, which explains most of source data dispersion (table 2), corresponds to an eigenvector with coordinates close to 0.50 (table 1). This means that for PC1 you can take the average of all the source images. When visualized, the position of the road is unambiguously determined, as well as sections with different types of forest vegetation. The interpretation result is similar to the visual analysis of the source images (figure 2) and the distribution maps of vegetation indices (figure 3b). When comparing PC1 with historical and cartographic documents (figure 1b), a deviation of the road contour is no more than 2 pixels – 10% of its maximum width.

The second principal component of PC2, which accounts for 14% of the total dispersion of the source data (table 2), is more dependent on the spectral brightness in R (table 1). Therefore, when visualized, there is a division only into the open soil (maximum values) and vegetation (minimum values). The vegetation cover is almost homogeneous, without a separation into deciduous and coniferous forests (for reference, figure 2, image in NIR).

On the third principal component of PC3, which accounts for 3% of the total dispersion of the source data, green phytomass regions with the maximum spectral brightness in NIR were revealed (tables 1, 2). The resulting image has features similar to the distribution maps of the vegetation indices NDVI and RVI (figure 3b).
Figure 2. Multispectral photography result.

Figure 3. Spectral reflectance: a) spectral curves of natural objects: 1 – deciduous forest, 2 – coniferous forest, 3 – the soil [10, simplified figure 4]; b) vegetation indices.
Figure 4. Principal component visualization.

Table 1. Eigenvectors of the correlation matrix.

|        | PC1 | PC2 | PC3 | PC4 |
|--------|-----|-----|-----|-----|
| G      | 0.53| 0.01| -0.45| 0.70|
| R      | 0.41| 0.86| 0.18| -0.21|
| RE     | 0.52| -0.32| -0.42| -0.66|
| NIR    | 0.51| -0.38| 0.75| 0.10|

Table 2. Eigenvalues of the correlation matrix.

| Eigenvalue | PC1 | PC2 | PC3 | PC4 |
|------------|-----|-----|-----|-----|
| %          | 3.25| 0.57| 0.10| 0.06|
| %          | 81% | 14% | 3%  | 2%  |

The fourth principal component of PC4 has a positive correlation with the spectral brightness values in G and a negative correlation with RE (table 1). This dependence determines the areas of young vegetation. It is during the early growing season that the spectral brightness increases in G, and
the minimum values in RE do not separate vegetation into coniferous and deciduous forests. When visualizing PC4, the following areas correspond to the areas of young vegetation (figure 5): 1 – trees planted in a row at the cutover area, 2 and 3 – ruderal vegetation along the roads, 4 – traces of the overgrown arable land. If regions 1–3 are unambiguously interpreted according to the image in the visible range, then region 4, as a site of human impact, is represented by historical cartographic documents (figure 1a). While doing so, the southern part of the field of the early 19th century has been revealed – about 67% of the arable land once used by the inhabitants of the Pudemsky Ironworks village.

![Figure 5](image.png)

Figure 5. Identified areas of young vegetation of the fourth principal component.

4. Conclusion
The proposed approach for the analysis of multispectral aerial photography data, based on the application of the PCA method, is effective in searching for areas of anthropogenic impact in historical territories. Typically, these areas do not have contrasting distinctive features that are evident in the original multispectral aerial images and the distribution maps of vegetation indices. It has been shown that traces of anthropogenic impact can be detected by visual analysis of the last principal components. This is where minor changes in the vegetation cover are registered that can be associated with areas of anthropogenic impact.

The applicability of the proposed approach is demonstrated by a case study of historical territory – the agricultural district of the Pudemsky Ironworks of the 18th – 19th centuries. An analysis of principle components determined the most informative channels in the early growing season – G and RE, which made it possible to reveal 67% of the arable land abandoned in the late 19th century and nowadays covered with forest vegetation. The results obtained are confirmed by historical cartographic documents. This approach, based on the detection of hidden patterns in the spectral brightness of vegetation, can be applied to studies of other historical territories.

Acknowledgments
This study was supported by the Russian Science Foundation. Project No 19-18-00322 “Comparative historical research of anthropogenic landscapes in different regions by means of unmanned aerial vehicles (the Tambov region and Udmurtia, mid-XVIII - early XX centuries)”.

References
[1] Woolf A 2016 *The Impact of Technology in History and Archaeology* ed J Benefield and A Robbins (Oxford: Raintree) p 56
[2] Verhoeven G J 2012 Near-infrared aerial crop mark archaeology: from its historical use to current digital implementations J. Arhaeol. Method Th. 19 pp 132–60

[3] Gennaro A, Candiano A, Fargione G, Mangiameli M and Mussumeci G 2019 Multispectral remote sensing for post-dictive analysis of archaeological remains. A case study from Bronte (Sicily) Archaeol. Prospection 26 pp 299–311

[4] Brooke C and Clutterbuck B 2020 Mapping heterogeneous buried archaeological features using multisensor data from unmanned aerial vehicles Remote Sens. 12 41

[5] He H, Liang T and Yu X 2016 Remote sensing clustering analysis based on object-based interval modeling Comput. Geosci. 94 pp 131–9

[6] Caspari G and Jedryke M 2017 Archsphere – A cluster algorithm for archeological applications J. Archaeol. Sci. Rep. 14 pp 181–8

[7] Chyla J M 2017 How can remote sensing help in detecting the threats to archeological sites in Upper Egypt? Geosci. 7 pp 63–74

[8] Jolliffe I T 2002 Principal Components Analysis. 2nd ed. (NewYork: Springer-Verlag New York Inc.) p 487

[9] Xue J and Baofeng S 2017 Significant remote sensing vegetation indices: a review of developments and applications J. Sens. 2017 p 17

[10] Pan P, Chen G, Saruta K and Terata Yu 2015 Snow cover detection based on visible red and blue channel from MODIS imagery data Int. J. of Geosci. 6 pp 51–66