Modeling boundaries of the occupation layer of an archaeological site on the basis of principal component analysis

A G Zlobina
Udmurt Federal Research Center of the Ural Branch of the Russian Academy of Sciences, 34 T. Baramzina str., Izhevsk, 426067, Russia

E-mail: Elf54@yandex.ru

Abstract. An approach is proposed for the integrated processing of the data of multispectral aerial photography and a geophysical survey; this approach is based on the consistent application of the principal component analysis and the fuzzy c-means clustering algorithm. Its use provides the identification of thick humus layer areas, clarification of the boundaries of a thick occupation layer, as well as the localization of sites presumably containing buried archaeological objects. This approach was successfully applied in identifying an occupation layer and modeling its boundaries in the test area of the archaeological site – Kushmanskoye fortified settlement Uchkakar (Udmurt Republic, Yarsky district).

1. Introduction and statement of the problem

The identification of a cultural layer and the determination of its boundaries is the initial stage in the study of an archaeological site. The solution to this problem is significantly complicated by the lack of topographic features of ancient structures, which are “smoothed out” as a result of intensive agricultural activity. Large-scale excavations are a traditional way of revealing an occupation layer, which leads to the destruction of the object of historical and cultural heritage. An alternative way to study historical territories is a comprehensive analysis of data obtained by multispectral aerial photography and a geophysical survey.

Multispectral aerial photography reveals patterns of changes in the vegetation cover. Areas of a thick humus layer formed as a result of anthropogenic impact (the occupation layer of an archaeological site) are characterized by higher vegetation intensity. When conducting a geophysical survey by means of an electrical survey, the distribution of soil resistivity is being analyzed. A thick occupation layer has a higher level of resistivity, which enables its detection. In either case, the identification of an occupation layer is achieved through the segmentation of primary data [1]. As a rule, clustering algorithms are used for this purpose [2–14].

The paper proposes a different approach to the identification of an occupation layer. This approach involves the consistent application of principal component analysis and the fuzzy c-means clustering algorithm. Principal component analysis is carried out in order to reduce the dimensionality of the initial data containing heterogeneous information about the study site: surface images in different spectral bands and maps of the distribution of soil resistivity. Thus, it is suggested to switch from clustering a large set of data arrays to segmenting the principal components having the greatest
dispersion. Segmentation is carried out by the fuzzy c-means clustering algorithm, which allows modeling the boundaries of the occupation layer. The choice of this algorithm for segmentation is justified by the fact that, when reconstructing the boundary of an object according to geophysical data, it proved to be more efficient than thresholding, region growing, gradient methods, hard clustering methods (k-means), and Bayesian classification methods (EM algorithm) [15]. The proposed approach to processing data of multispectral aerial photography and a geophysical survey has been used to study the medieval archaeological site – Kushmanskoye settlement Uchkakar (Udmurt Republic, Yarsky district).

2. Data processing methods

2.1. Principal component analysis
The basic idea of principal component analysis (PCA) is to isolate groups of closely correlated variables in primary data space \( X = (x_1, x_2, ..., x_n) \) and to replace them with principal components \( Y = (y_1, y_2, ..., y_m) \) without loss of any informational value.

PCA data conversion is performed according to the following algorithm [16]:

1. Standardization of primary data. Thus, variables have unit variance (the same information content) and the origin is transferred to the center of the data cloud. Further, for the sake of simplicity, we understand standardized primary data.

2. Calculation of eigenvectors and eigenvalues \( (\lambda_1, \lambda_2, ..., \lambda_n) \) of the correlation matrix of the primary data. Eigenvalues are sorted in descending order: \( \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \).

3. Calculation of the principal components:
\[
y_j = \sum_{i=1}^{n} \alpha_{ij} \cdot x_i, \quad j = 1, ..., n,
\]
where \( \alpha_{ij} \) is the correlation matrix eigenvector corresponding to \( \lambda_j \). With \( \lambda_j = D(y_j) \), the calculated principal components will also be ordered (in descending order of variance).

4. The reduction of dimension \( Y = (y_1, y_2, ..., y_m) \) due to the exclusion of uninformative variables.

To solve this problem, we apply the Kaiser criterion [17], according to which it is sufficient to consider only the principal components \( y_i \) \( (i = 1, ..., m, m < n) \) that eigenvalues \( \lambda_i \geq 1 \) correspond to.

The application of PCA is carried out not only to reduce the dimensionality of the primary data, but also to detect hidden patterns that are not detected in the analysis of individual variables \( x_i \).

2.2. Fuzzy c-means algorithm
The application of the fuzzy c-means algorithm implies the partition of the primary data \( y_k = (y_{k1}, y_{k2}, ..., y_{km}) \) into a given number \( c \) \( (2 \leq c \leq n) \) of fuzzy classes. And fuzzy classes are governed by a fuzzy partition matrix:
\[
U = [\mu_{ij}], \quad \mu_{ij} \in [0,1], \quad \sum_{j=1}^{c} \mu_{ij} = 1, \quad \sum_{i=1}^{n} \mu_{ij} > 0, \quad i = 1, ..., n, \quad j = 1, ..., c,
\]
where \( \mu_{ij} \) – the degree to which element \( y_{ki} \) belongs to class \( j \).

The fuzzy c-means algorithm is iterated according to the following scheme [18]:

1. Initialization. Input parameters are set: \( c \) – the number of classes, \( m \) – exponential weight of fuzzy clustering (usually, \( m = 2 \)), \( \varepsilon \) – breakpoint parameter. The fuzzy partition matrix \( U \) is set randomly.

2. Calculation of the centers of fuzzy classes:
3. Recomputation of the elements of a fuzzy partition matrix:

\[
\mu_{ij} = \left( \sum_{g=1}^{c} \left( \frac{\|y_{ki} - v_{g}\|^2}{m-1} \right) \right)^{-1}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, c.
\]

If \(\|y_{ki} - v_{g}\| = 0\) is fulfilled for a certain class \(g\), then we set \(\mu_{ig} = 1\), and \(\forall j \neq g \quad \mu_{ij} = 0\).

4. Checking the optimality of the current partition:

\[
\|U - U^*\| < \varepsilon,
\]

where \(U^*\) – a fuzzy partition matrix at the previous iteration. If the condition is fulfilled, then the algorithm terminates; otherwise, go to step 2.

The result of the algorithm is a finite fuzzy partition matrix. An element is assigned to the class attribution to which is stronger than that to other classes.

3. Field experiment, its result and discussion

A field experiment was performed on the territory of the Kushman settlement Uchkakar (Udmurt Republic, Yarsky district). This archaeological site, dated to the 9th–13th centuries, is located on the right bank of the Cheptsa River on a high cape of a subtriangular shape. The steep slopes in the south, north, and northwest, clearly visible on the orthophotomap (figure 1a), determine the probable boundaries of the settlement. Topographically, only the vague foundations of two lines of fortification (the middle and exterior banks) are visible. Throughout the rest of the settlement territory, the surface is almost flat; the evidence of medieval buildings is indiscernible. Selective excavations and soil drilling were carried out on the cape of the settlement (limited by the middle bank), which proved the presence of the cultural layer, revealed trends in a change of its thickness and discovered the remains of diverse archaeological sites (fortification, pit-houses, clay foundations of buildings, etc.) [19]. It was the diversity of the structure and condition of the cultural layer that determined the choice of this site for testing the proposed approach. The shape and boundaries of the test site (red outline in figure 1a) are determined by the landscape boundaries of the territory studied and archaeological features of topography.

The territory of the settlement has been subjected to multispectral aerial photography, which has yielded images in green (0.52–0.60 microns), red (0.63–0.69 microns) and near infrared (0.75–0.90 microns) spectral bands. When analyzing an image in the visible spectral band (figure 1a), only footpaths, an excavation and a site adjacent to it can be unambiguously identified on the test site. These areas of ruderal vegetation can be clearly seen in the images in the red and near infrared spectral bands: the red spectral band has the maximum absorption of solar radiation by chlorophyll (figure 2a); whereas the near infrared band has the maximum reflection (figure 2b) [20]. However, their contrast accentuation is caused not by the presence of the occupation layer of the archaeological site, but rather by modern anthropogenic impact on the study area.
Figure 1. Kushmanskoye fortified settlement Uchkakar: (a) orthophotomap, aerial photography in the visible spectrum, (b) diagram of the study site (1 – boundaries of the test site, 2 – boundaries of the excavation).
Figure 2. Visualization of a multispectral image: (a) the red spectral band, (b) the near infrared spectral band.

Since an analysis of the images provided no information about the presence of the occupation layer, the initial data were converted by the PCA method. When visualizing the first principal component (figure 3a, $\lambda_1 = 1.9$), only an excavation, having rectangular outlines, and several bushes have been clearly distinguished. However, the segmentation of this component (figure 3b) has showed a change in the thickness of the humus layer (classes I, II, III), thereby dividing the test section into two fundamentally different regions – R1 and R2 (figure 1b). The region R1 as a whole is characterized by a thick humus layer (class III) with inclusions of thinner local sections (class II). The structure, which is heterogeneous in humus thickness, can be bind together due to the presence of archaeological objects (pit-houses, clay foundations of buildings, etc.) buried in the soil. A different structure was revealed in region R2. It is characterized by the evident predominance of thinner humus layers (classes I, II). In this case, there is an increase in the humus thickness in the northeast direction – towards the region R1.

The origin of the identified humus layer can be caused by both anthropogenic impact (occupation layer of the archaeological site) and natural factors (gradual accumulation of the soil layer in negative landforms). Thus, further study of the test site is necessary, one aimed at assessing the thickness of the humus layer.

It is possible to assess the thickness of a possible occupation layer by conducting a geophysical survey, using the method of multispatial electric profiling. Based on the results of such a survey, three data arrays were formed – distribution maps of apparent resistivity $\rho_k$ at effective depths of 0.32 m, 0.45 m, and 0.68 m. Since the occupation layer has an increased level of resistivity, it will be manifested as areas of high $\rho_k$ values. Visualization of the first principal component (figure 4a, $\lambda_1 = 2.8$), obtained by applying the PCA method, provided an average representation of the thickness of the occupation layer over the three arrays under consideration. In this case, the division of the test site into the R1 and R2 regions (figure 1b) is more contrasting. After segmenting the first principal component (figure 4b), the region R1 almost completely corresponds to the thick occupation layer (class III). The thickness of the occupation layer gradually decreases at the boundaries of the test site, located along the hillside (class I, II). Such a situation may be the result of erosion processes caused by long-term plowing of the settlement territory. Region R2, on the contrary, is characterized by a predominance of the thinnest occupation layer (class I).
Figure 3. Processing multispectral aerial photography data: (a) the first principal component, (b) the result of segmentation (1 – class I, 2 – class II, 3 – class III, 4 – excavation boundaries).

Figure 4. Processing geophysical data: (a) the first principal component, (b) the result of segmentation (1 – class I, 2 – class II, 3 – class III, 4 – the boundaries of the bank and the ditch).

It should also be noted that only a geophysical survey managed to reveal the internal defense line of the settlement (bank and ditch). This line of defense, which has been revealed during excavations and has no topographic features, limits from the southwest the area of a thick occupation layer. Thus, the outer boundaries of the ditch can be regarded as the dividing line between the regions R1 and R2 in figure 1b.
A similar division of the test area into regions R1 and R2 is observed with PCA processing of the entire data set (figure 5a, $\lambda_1 = 3.7$), obtained as a result of the complex analysis of distribution maps $\rho_k$ and surface images in different spectral bands. The segmentation of the first principal component (figure 5b) also made it possible to evaluate the distribution of the thickness of the occupation layer (classes I, II, III). But unlike when processing geophysical data (figure 4b), a group of local anomalous areas belonging to class II was revealed in the region of a thick cultural layer (class III). Their occurrence is likely due to archaeological objects – presumably sites of compacted and calcined clay [19] – being buried in the ground. Such clay platforms were the central part of medieval residential, household, and industrial buildings. They lie directly under the layer of turf; and the clay mass itself, which has less resistivity, spreads in depth down to the mainland layer. For this reason, vegetation intensity changes and the level of soil resistivity in the corresponding local areas decreases. Therefore, the identification of such archaeological objects is possible only due to the integrated processing of data of multispectral aerial photography and a geophysical survey.

The obtained information about the probable location of archaeological sites buried in the ground (in this particular case, the clay foundations of buildings) can be used later, when planning targeted excavations. This will significantly reduce the labor costs of exploring an archaeological site.

4. Conclusion

The proposed approach to processing data of multispectral aerial photography and a geophysical survey, involving the consistent application of the PCA method and the segmentation of the first principal components by the fuzzy c-means algorithm, is effective in identifying the occupation layer of an archaeological site. The study is conducted on a test site (Kushmanskoye fortified settlement Uchkakar) where there are no topographic signs of evidence for ancient structures. It is shown that due to processing data of multispectral photography on a site with a relatively flat surface, it is possible to localize areas of a thick humus layer (the occupation layer of an archaeological site). This justifies the need for further research. Processing data of multispectral aerial photography and a geophysical survey makes it possible to evaluate the thickness of the cultural layer and, consequently, to model its
boundaries. The integrated data processing of multispectral aerial photography and a geophysical survey provides additional information on the layout of the archeological site: anomalous areas identified within a thick occupation layer are most likely to be associated with buried archaeological sites.

Therefore, the proposed approach to processing data of multispectral aerial photography and a geophysical survey enables not only the identification of the cultural layer and determination of its boundaries, but also planning for further archaeological work aimed at searching for middle-age buildings.

Acknowledgments
The study was carried out with the financial support of the Russian Federal Property Fund (project No. 18-49-180007).

References
[1] Gonzalez R C and Woods R E 2002 Digital image processing, 2nd ed. (New Jork: Rrentice Hall) p 793
[2] Rowlands A and Sarris A 2007 Detection of exposed and subsurface archaeological remains using multi-sensor remote sensing J. Archaeol. Sci. 34 pp 795–803
[3] Song Yu, Meng H, O’Grady M J and O’Hare G. M. P. 2010 The application of cluster analysis in geophysical data interpretation Computat. Geosci. 14 (2) pp 263–71
[4] Doneus M, Verhoeven G, Atzberger C, Wess M and Rus M 2014 New ways to extract archeological information from hyperspectral pixels J. Archaeol. Sci. 52 pp 84–96
[5] Giuseppe M G, Troiano A, Troise C and Natale G 2010 The application of cluster analysis in geophysical data interpretation Comput. Geosci. 14 (2) pp 263–71
[6] He H, Liang T and Yu X 2016 Remote sensing clustering analysis based on object-based interval modeling Comput. Geosci. 94 pp 131–9
[7] Caspari G and Jedryke M 2017 Archsphere – A cluster algorithm for archeological applications J. Archaeol. Sci. Rep. 14 pp 181–8
[8] Chyla J M 2017 How can remote sensing help in detecting the threats to archeological sites in Upper Egypt? Geosci. 7 (4) 63–74
[9] Deiana R, Leucci G and Martorana R 2018 New perspectives on Geophysics for Archaeology: A Special Issue Surv. Geophys. 39 pp 1035–8
[10] Zlobina A G, Zhurbin I V and Nemtsova O M 2018 Selective segmentation of the shallow electrical profiling data Optoelectron. Instrum. Data Proces. 54 pp 340–7
[11] Scudero S, Martorana R, Capizi P, Pisciotta A, D’Alessandro A, Bottari C and Di Stefano G 2018 Integrated geophysical investigations at the Greek Kamarina site (Southern Sicily, Italy) Surv. Geophys. 39 pp 1181–200
[12] Davis D S, Lipo C P and Sanger M C 2019 A comparison of automated object extraction methods for mound and shellring identification in coastal South Carolina J. Archaeol. Sci. Rep. 23 pp 166–77
[13] Khalaf N and Insoll T 2019 Monitoring islamic archaeological landscapes in Ethiopia using open source satellite imagery J. Field Archaeol. 44 pp 401–19
[14] Thabeng O L, Merlo S and Adam E 2019 High-resolution remote sensing and advanced classification techniques for the prospection of archeological sites’ markers: The case of 7dung deposits in the Shashi-Limpopo Confluence area (southern Africa) J. Archaeol. Sci. 102 pp 48–60
[15] Zlobina A G and Zhurbin I V 2018 Comparative analysis of segmentation algorithms for electrical profiling data when reconstructing the boundaries of an object Inf. Technol. 24 pp 805–12 [in Russian]
[16] Jolliffe I T 2002 Principal Components Analysis. 2nd ed. (New York: Springer-Verlag New
York Inc.) p 487

[17] Kaiser H F 1960 The application of electronic computers to factor analysis Educ. Psychol. Meas. 20 pp 141–51

[18] Bezdek J C 1981 Pattern Recognition with Fuzzy Objective Function Algorithms (New York: Plenum Press) p 272

[19] Zhurbin I V and Ivanova M G 2018 Geophysical Studies at the Kushmanskoye (Uchkakar) Fortified Settlement, Kama Basin Archaeol. Ethnol. Anthropol. Eurasia 46 pp 76–85

[20] Schowengerdt R A 2007 Remote sensing. Models and methods for image processing. 3rd ed. (Amsterdam: Elsevier) p 560