PRACE INSTYTUTU LOTNICTWA ISSN 0509-6669
Nr 1 (234), s. 143-150, Warszawa 2014

ANALIZA CZYNNIKOWA ZDJĘĆ WIELOSPEKTRALNYCH

CZAPSKI PAWEŁ*, KOTLARZ JAN*, KUBIAK KATARZYNA**, TKACZYK MIŁOSZ**

** Instytut Lotnictwa, Zakład Teledetekcji Centrum Technologii Kosmicznych*

*** Instytut Badawczy Leśnictwa, Zakład Ochrony Lasu*

Streszczenie

Analiza zdjęć wielospektralnych sprowadza się często do modelowania matematycznego opartego o wielowymiarowe przestrzenie metryczne, w których umieszcza się pozyskane za pomocą sensorów dane. Tego typu bardzo intuicyjne, łatwe do zaaplikowania w algorytmice analizy obrazu postępowanie może skutkować zupełnie niepotrzebnym wzrostem niezbędnej do analiz zdjęć mocy obliczeniowej. Jedną z ogólnie przyjętych grup metod analizy zbiorów danych tego typu są metody analizy czynnikowej. W pracy tej prezentujemy dwie z nich: Principal Component Analysis (PCA) oraz Simplex Shrink-Wrapping (SSW). Użyte jednocześnie obniżają znacząco wymiar zadanej przestrzeni metrycznej pozwalając odnaleźć w danych wielospektralnych charakterystyczne składowe, czyli przeprowadzić cały proces detekcji fotografowanych obiektów. W roku 2014 w Pracowni Przetwarzania Danych Instytutu Lotnictwa oraz Zakładzie Ochrony Lasu Instytutu Badawczego Leśnictwa metodykę tą równie skutecznie przyjęto dla analizy dwóch niezwykle różnych serii zdjęć wielospektralnych: detekcji głównych składowych powierzchni Marsa (na podstawie zdjęć wielospektralnych pozyskanych w ramach misji EPOXI, NASA) oraz oszacowania bioróżnorodności jednej z leśnych powierzchni badawczych projektu HESOFF.

Słowa kluczowe: PCA, metody statystyczne, bioróżnorodność, krzywe blasku, redukcja danych

MATEMATYCZNE PODSTAWY REDUKCJI ZBIORU DANYCH

Obrazując dowolny obiekt poprzez zdjęcia wielospektralne otrzymujemy informację o natężeniu światła emitowanego lub odbitego przez ten obiekt w przynajmniej kilku długościach fali elektromagnetycznej. Informację taką możemy zapisać jako punkt I w n-wymiarowej przestrzeni metrycznej, gdzie n jest liczbą długości fal elektromagnetycznych (kolorów) w których rejestrujemy promieniowanie (fotografujemy obiekt)

$$I = (I(\lambda_1), I(\lambda_2), I(\lambda_3), \ldots, I(\lambda_n)).$$

Jeśli w trakcie pozyskiwania danych prowadzimy obserwację zjawiska w czasie (np. rejestracja światła odbitego od powierzchni dalekiej planety) lub obrazujemy dany obiekt przestrzennie (np. wykonujemy zdjęcia kompleksów leśnych) takich punktów w naszym zbiorze

obserwacji będzie bardzo wiele. Należy tutaj zauważyć, że rejestrowane tutaj natężenia światła docierającego od badanego obiektu w różnych długościach fali będą zależeć przede wszystkim od tego jakie materiały (związki chemiczne, minerały, typy powierzchni, gatunki biologiczne itp.) składają się na obrazowany obiekt lub obszar. Można przypuszczać, że dla bardzo licznych zbiorów danych obrazujące owe natężenia punkty będą skupione wokół punktów obrazujących sygnatury spektralne komponentów odwzorowane na wybrane długości fal elektromagnetycznych, albo będą tych punktów kompozycją.

W celu określenia samej ilości zobrazowanych obiektów z pomocą przychodzą metody analizy czynnikowej, a w szczególności analiza głównych składowych (Principal Component Analysis).

1.1. Principal Component Analysis

Załóżmy, że w ramach obrazowania wielospektralnego pozyskano zdjęcia w n długościach fali elektromagnetycznej. Zebrane dane opisujemy zatem jako punkty w n-wymiarowej przestrzeni liniowej, gdzie każdemu wymiarowi odpowiada natężenie światła docierającego od fotografowanego obiektu w fali o długości λ_n. Celem analizy czynnikowej jest zredukowanie dużej liczby zmiennych do prostszego w opisie zbioru. W przypadku PCA będzie to znaczna redukcja liczby wymiarów przestrzeni przy bardzo niewielkiej utracie dokładności zgromadzonych danych.

W efekcie zastosowania algorytmu opartego o PCA obrócone zostają osie układu współrzędnych w taki sposób, aby w pierwszej kolejności zmaksymalizować wariancję pierwszej współrzędnej, następnie drugiej współrzędnej itd. Przyjmuje się, że analizę problemu możemy zredukować do co najmniej takiej ilości wymiarów przestrzeni, by suma odpowiadających im wariancji stanowiła co najmniej 99% sumy wariancji danych we wszystkich wymiarach. Warto tutaj zauważyć, że jeśli fotografowany obiekt zawierał i różnych elementów składowych o różnych sygnaturach spektralnych, to redukcja problemu powinna nastąpić do i-1 wymiarów [3].

1.2. Simplex Shrink-Wrapping

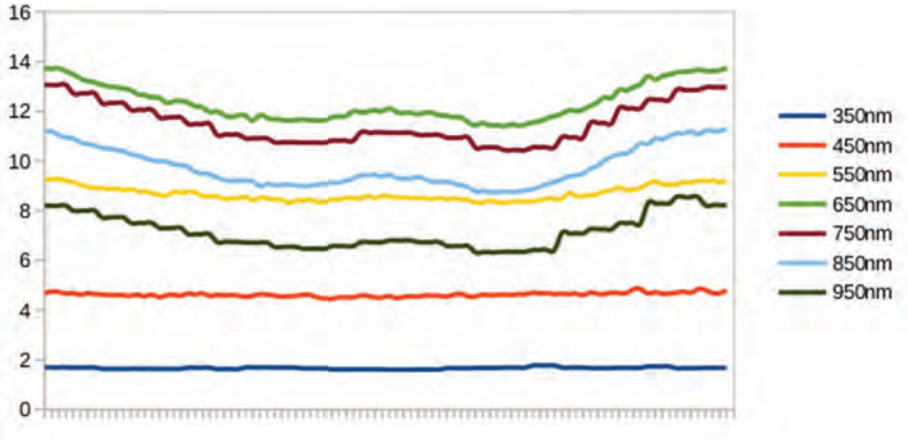
Algorytmika PCA upraszacza analizę przede wszystkim poprzez redukcję wymiaru przestrzeni danych obserwacyjnych. Zrzutowanie danych na i osi sugeruje zatem obecność na fotografowanym obszarze $i+1$ komponentów o różnych sygnaturach spektralnych. Należy zauważyć, że umieszczone w i-wymiarowej przestrzeni danych $i+1$ punktów będzie tworzyło i-wymiarowy simpleks. Zauważmy, że aby otrzymane w rezultacie sygnatury spektralne poszczególnych komponentów miały sens, to sympleks ten powinien zawierać wszystkie punkty reprezentujące dane obserwacyjne, a z wielu sympleksów o tej właściwości powinniśmy wybrać ten o najmniejszej *mierze głównej*. Wierzchołki simpleksu o tych dwóch właściwościach reprezentują w efekcie sygnatury spektralne komponentów.

Jednym z algorytmów znajdujących taki simpleks jest „Simplex Shrink-Wrapping"[4].

2. PROBLEM DETEKCJI PODSTAWOWYCH TYPÓW POWIERZCHNI MARSA

Zmienna jasność Marsa jest obserwowana od wielu dziesięcioleci. Jednym z głównych czynników wpływających na tą zmienność w krótkim, liczonym w godzinach, czasie jest obrót planety wokół własnej osi. Zmienność ta jest szczególnie łatwa do zaobserwowania podczas wykonywania zdjęć wielospektralnych, głównie w zakresie 650 – 1000 nm. Dane obserwacyjne na podstawie których podjęta została próba detekcji typów powierzchni planety pochodzą

z misji EPOXI i zostały zarejestrowane w czasie od 2009-11-20 11:53:23.987 do 2009-11-21 11:53:21.678 UTC, zatem zawierają czas w którym Mars wykonuje jeden pełen obrót wokół własnej osi. Natężenie światła docierającego z planety do detektorów EPOXI ilustruje wykres 1.



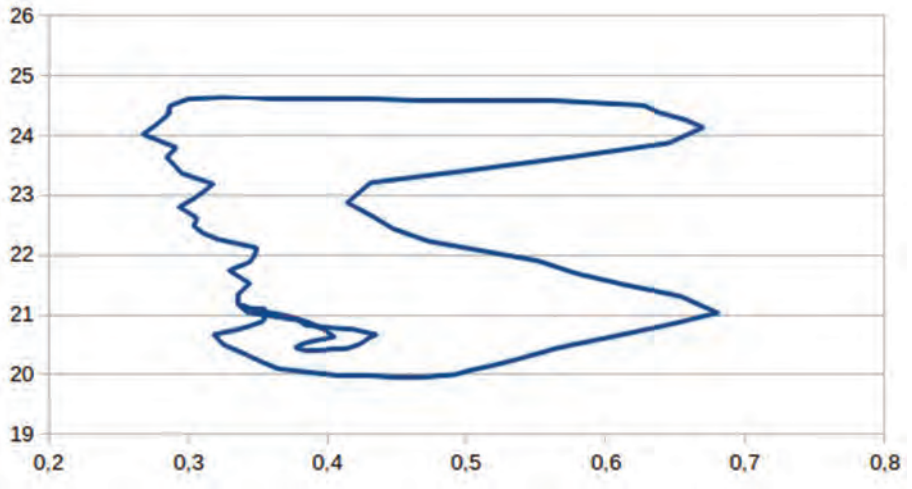
Wykres 1: Zależność natężenia światła docierającego do sensorów EPOXI (W/m^2 * 10^4) od czasu w różnych długościach fali elektromagnetycznej. Dla poszczególnych filtrów optycznych w trakcie obrotu planety obserwujemy różną zmienność natężenia światła: minimalną dla fal UV (350 nm), maksymalną dla koloru czerwonego i podczerwieni (650 - 850 nm).

W ramach eksperymentu zarejestrowano światło w siedmiu długościach fali od 350 do 950 nm (UV, promieniowanie widzialne, bliska podczerwień). Zgodnie z przyjętą metodyką umieszczamy zaobserwowane dane w siedmiowymiarowej przestrzeni liniowej i poddajemy je analizie algorytmem PCA. W efekcie otrzymujemy nową, dwuwymiarową przestrzeń, na którą rzutujemy dane obserwacyjne. Ortogonalne wektory, które rozpinają nową przestrzeń zdefiniowane w dotychczasowej przestrzeni siedmiowymiarowej otrzymujemy z algorytmu PCA:

e_1 = (0.0426102881, 0.0105459055, -0.0198743574, -0.0485914897, -0.7083607763, 0.551320319, 0.4354169734)
e_2 = (-0.0428064746, -0.2724816615, 0.3896674024, -0.5221739046, 0.4579646004, 0.1588109247, 0.514259267)

Wariancja danych wzdłuż wektora e_1 stanowi 98,5% całkowitej wariancji danych. Wariancja wzdłuż wektora e_2 stanowi nieco ponad 0,5% całkowitej wariancji. Sumaryczna wariancja na otrzymanej w wyniku zastosowania algorytmu PCA wariancja wynosi zatem powyżej 99% całkowitej wariancji. Otrzymany wynik mówi nam, że zmienność Marsa w trakcie obrotu przebiega głównie w trzech pasmach optycznych opisujących bliską podczerwień: 750, 850 i 950 nm (w tych trzech wymiarach przestrzeni siedmiowymiarowej rozpina się głównie wektor e_1). Drugi z wektorów mówi nam o zmienności w zakresach barw: zielonej, czerwonej i podczerwieni, jednak ta zmienność jest niemal dwustukrotnie mniejsza od tej wyrażonej tylko w trzech kanałach podczerwieni.

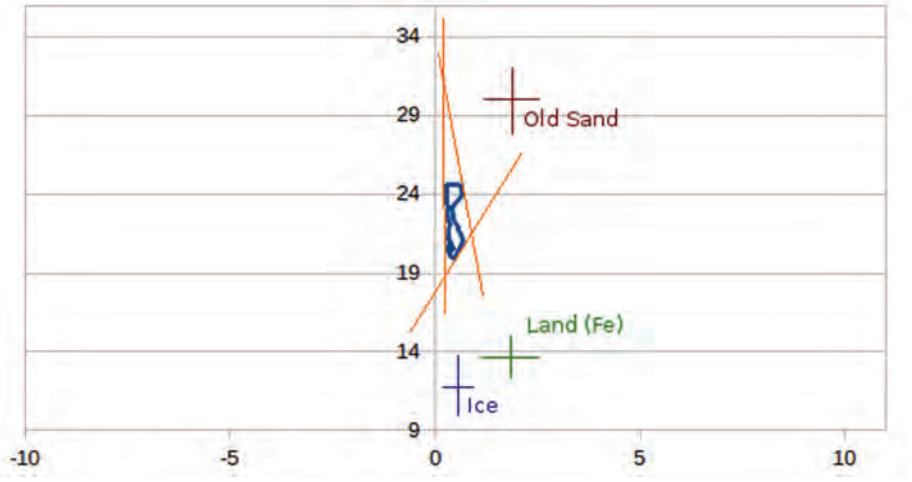
Po zrzutowaniu zbioru danych na przestrzeń rozpiętą wektorami e_1 i e_2 otrzymujemy zbiór punktów przygotowanych do analizy algorytmem SSW.



Wykres 2: Dane obserwacyjne zrzutowane na przestrzeń dwuwymiarową rozpiętą na wektorach e_1 (oś pionowa) i e_2 (oś pozioma).

Ponieważ rzutowanie danych obserwacyjnych nastąpiło na przestrzeń dwuwymiarową ($i = 2$) simpleksem będzie tutaj trójkąt, a miarą główną będzie pole powierzchni.

Po analizie algorytmem SSW otrzymujemy szukany simpleks wraz ze współrzędnymi jego wierzchołków:



Wykres 3: Trójkąt zawierający wszystkie punkty reprezentujące dane pomiarowe o najmniejszej powierzchni wraz z naniesionymi punktami reprezentującymi glebę o dużej zawartości żelaza, lód wodny oraz piasek (kwarcowy).

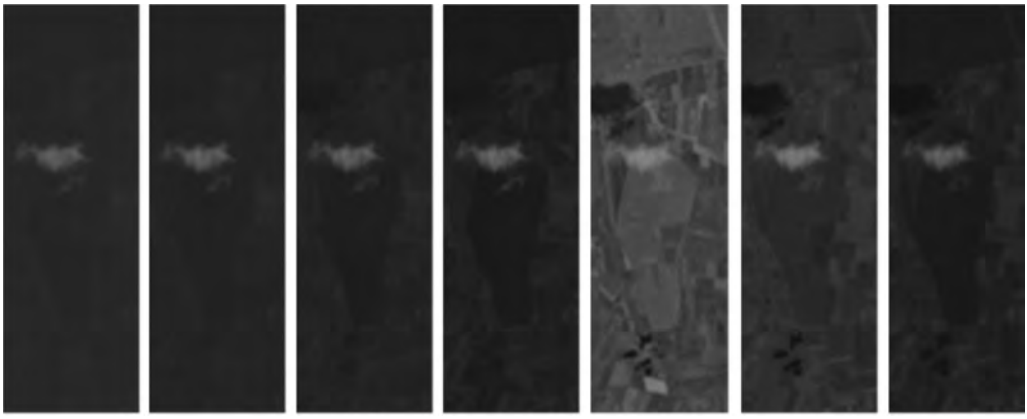
Trzy punkty wierzchołkowe definiujące simpleks próbujemy interpretować jako rzuty sygnatur spektralnych konkretnych materiałów na naszą powierzchnię. Na wykresie 3 podane zostały trzy prawdopodobne materiały: lód wodny, gleba bazaltowa z dużą domieszką żelaza oraz piasek (kwarc). Widzimy, że punkty te położone są relatywnie daleko od wykresu obrazującego ewolucję planety na kolejnych zdjęciach, jednak trzeba zauważyć, że ponieważ planeta

zobrazowana została tutaj jako jeden piksel, to na każdą z danych cząstkowych składały się dane aż z połowy planety. Oczywiście można tutaj poszukiwać innych sygnatur spektralnych, bardziej odpowiadających wierzchołkom simpleksu, jednak jest to już kwestia interpretacji wyników a nie samej algorytmiki.

OPIS BIORÓŻNORODNOŚCI POWIERZCHNI BADAWCZEJ PROJEKTU HESOFF

Zupełnie innym zbiorem danych dla którego możemy zastosować przedstawioną powyżej algorytmikę jest zbiór danych przestrzennych takich jak zdjęcie lotnicze lub satelitarne. O ile jednak dla danych rejestrowanych w czasie w rezultacie otrzymujemy samą informację o możliwych składowych sygnaturach spektralnych badanego obiektu, o tyle w przypadku danych typu przestrzennego pozyskaną informację możemy zaaplikować do stworzenia orientacyjnej mapy zdefiniowanych komponentów.

Jako dane źródłowe w tym przypadku pobieramy zdjęcia satelitarne powierzchni badawczej realizowanego w Instytucie Lotnictwa i Instytucie Badawczym Leśnictwa projektu HESOFF (w nadleśnictwie Krotoszyn). Zdjęcie wykonane przez instrument OLI umieszczony na pokładzie satelity Landsat-8 pochodzi z lipca 2013 roku, a zatem z okresu wysokiego poziomu wegetacji drzew. Do dyspozycji otrzymujemy siedem zdjęć wykonanych w zakresie 350 – 800 nm (pasma UV i VIS).



Rys. 1. Powierzchnia badawcza projektu HESOFF w nadleśnictwie Krotoszyn zobrazowana na zdjęciach satelitarnych Landsat-8, lipiec 2013, pasmo 350 – 800 nm.

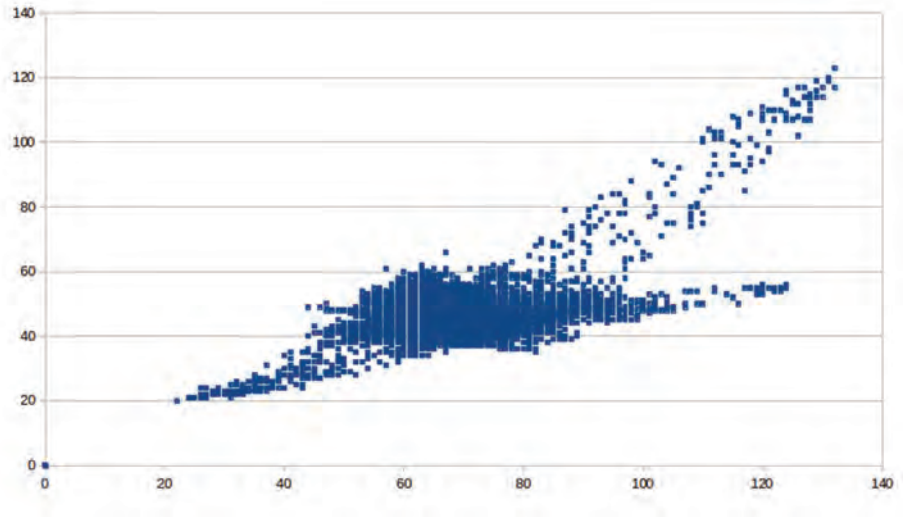
Piąta fotografia od lewej strony prezentuje reflektancję w barwie zielonej. Można na niej dostrzec, że cały kompleks leśny różnicuje się na dwie zasadnicze części. Jaśniejsze natężenie barwy zielonej otacza nieco ciemniejszą plamę (w połowie wysokości zdjęcia, od strony prawej). W tym przykładzie spróbujemy zdiagnozować przyczynę tego zróżnicowania.

Ponownie do dyspozycji mamy siedem długości fali elektromagnetycznej, zatem pierwotna przestrzeń będzie tak jak w przykładzie poprzednim tj. siedmiowymiarowa. Współrzędne poszczególnych punktów będziemy obliczać tym razem jednak w ten sposób, że dla kolejnych pikseli zdjęć będziemy szukać odpowiadających im pikseli na innych zdjęciach, a współrzędnymi będą wartości z zakresu (0,255) natężenia jasności pikseli w poszczególnych pasmach optycznych. Zobrazowanie powierzchni stanowią zdjęcia o rozdzielczości 51 x 151 pikseli. W sumie zatem punktów obrazujących dane będziemy mieć 7 701.

38;34;31;26;70;40;27;
38;34;31;26;71;41;27;
38;34;31;26;69;40;27;
38;34;31;26;64;39;27;
38;34;31;26;65;40;27;
38;34;31;26;69;41;28;
38;34;31;26;64;39;27;

Rys. 2. Fragment pliku zawierającego współrzędne poszczególnych pikseli w siedmiowymiarowej przestrzeni

Przygotowane w ten sposób dane poddajemy analizie algorytmem PCA. Jeśli ograniczymy się rzeczywiście do obszaru tylko leśnego (pominiemy pola oraz obszar widocznej na zdjęciu chmury) otrzymamy dwuwymiarową przestrzeń z sumą wartości własnych 99.6%. Podobnie jak poprzednio odwzorowujemy nasze dane na tą przestrzeń otrzymując następujący wynik:



Wykres 4: Rozkład pikseli ze zdjęcia satelitarnego na otrzymanej dwuwymiarowej przestrzeni głównych składowych.

Widzimy tu bardzo precyzyjny rozkład danych. Umieszczając te dane w algorytmie SSW otrzymalibyśmy trzy charakterystyczne sygnatury spektralne – typowe dla lasu liściastego, lasu iglastego oraz dla lasu mieszanego. Większość z pikseli znajduje się bardzo blisko sygnatur spektralnych lasu liściastego oraz lasu mieszanego. Stosunkowo niewiele z nich znajduje się najbliżej lasu iglastego.

W przeciwieństwie do poprzedniego przykładu dane te można teraz zrzutować na podkład mapowy:

Rys. 3: Naniesione na podkład mapowy oznaczenie natężenia występowania drzew iglastych. Im jaśniejsza barwa w ramach kompleksu leśnego tym większa ilość drzew iglastych. Poza kompleksem leśnym analiza jest błędna.

Na podstawie analizy algorytmami PCA oraz SSW wnioskujemy, że przyczyną zróżnicowania kompleksu leśnego na zdjęciach satelitarnych jest różny w poszczególnych pikselach udział drzew liściastych i iglastych.

WNIOSKI

Zastosowanie algorytmiki analizy czynnikowej wydatnie ułatwia zarówno proces obliczeń numerycznych podczas analizy zdjęć wielospektralnych jak również interpretację otrzymanych wyników. Główną wadą tego rozwiązania jest utrata zróżnicowania sygnału w niektórych pasmach optycznych. Jak pokazują oba przykłady strata ta sięga ok. 1% całej wariancji otrzymanego do analizy sygnału i dotyka głównie tych kanałów optycznych, które nie wnoszą znaczącego zróżnicowania do zarejestrowanych danych.

Kwestią otwartą pozostaje sposób interpretacji otrzymanych punktów wierzchołkowych simpleksów. Stosowane są tu dwa podejścia. Pierwsze z nich zakłada obliczenie możliwych sygnatur spektralnych wierzchołków. W tym podejściu największym problemem jest wysoki wymiar przekształcenia liniowego z przestrzeni wynikowej do przestrzeni pierwotnej (przekształcenie 7-2 = 5 wymiarowe w obu przykładach). Daje to ogromne zróżnicowanie możliwych sygnatur komponentów. Łatwiejsze do aplikacji wydaje się podejście drugie, gdzie nanosimy na przestrzeń wynikową punkty obrazujące sygnatury spektralne materiałów, które spodziewamy się odnaleźć na danej powierzchni (przykład 1).

LITERATURA

[1] Cowan N. B. et al., „DETERMINING REFLECTANCE SPECTRA OF SURFACES AND CLOUDS ON EXOPLANETS ", Astropysical Journal, 01/2013
[2] Cowan, N. B., et al. 2009, Astropysical Journal, 700, 915
[3] W. J. Krzanowski, „Principles of Multivariate Analysis: A User's Perspective", Oxford University Press, 2000

[4] Daniel R. Fuhrmann, „Simplex shrink-wrap algorithm", Proc. SPIE 3718, Automatic Target Recognition IX, 501 (August 24, 1999);
DOI:10.1117/12.359990; http://dx.doi.org/10.1117/12.359990

PRINCIPAL COMPONENT ANALYSIS OF MULTISPECTRAL IMAGES

Abstract

Mostly, analysis of multispectral images employs mathematical modeling based on multi-dimensional metric spaces that includes collected by the sensors data. Such an intuitive approach easily applicable to image analysis applications can result in unnecessary computing power increase required by this analysis. One of the groups of generally accepted methods of analysis of data sets are factor analysis methods. Two such factor analysis methods are presented in this paper, i.e. Principal Component Analysis (PCA) and Simplex Shrink - Wrapping (SSW). If they are used together dimensions of a metric space can be reduced significantly allowing characteristic components to be found in multispectral data, i.e. to carry out the whole detection process of investigated images. In 2014 such methodology was adopted by Data Processing Department of the Institute of Aviation and Division of Forest Protection of Forest Research Institute for the analysis of the two very different series of multispectral images: detection of major components of the Mars surface (based on multispectral images obtained from the epoxy mission, NASA) and biodiversity estimation of one of the investigated in the HESOFF project forest complexes.

Keywords: PCA, statistical methods, biodiversity, light curves, data reduction