Data Mining Methods in Vectorbased GIS[1]

Istvan Elek[2]

June 7, 2006

[1] Leopoldina Meeting, Thematic Mapping in Geosciences – Applications using New Technologies and Media, Budapest, May 25-27, 2006

[2] Eotvos Lorand University, http://lazarus.elte.hu/ ˜ elek

Abstract

The satellite image processing often uses data clustering methods to make images segmented. The result of the segmentation is a supervised or unsupervised classification on the certain image. In the vector based GIS there is a well known simple segmentation technique that is the thematic mapping. Unfortunately the vector GIS does not use clustering methods, although the thematic mapping technique takes only one kind of data into consideration. The classical thematic mapping neglects the most of data except for only one, but the classification would be much better probably if you take all data into account.

The method introduced in this article suggests a new approach for thematic mapping based on the built-in software solutions in the existing GIS software. The main concept of this approach is the application of principal component analysis which produces the first principal component being the target of thematic mapping.

1.1 Introduction

The thematic map is a kind of classification where the procedure takes one data column into consideration. In the most cases this result is proper. Graphic symbols, styles of classes are depend on the value of a certain data field. Sometimes the number of classes are fixed, but boundaries are not. In other cases the boundaries are pending regarding the classification algorithm.

Look at a simple example. Let us suppose we have a database on Hungarian settlements with 50 kind of attribute data (50 columns) such as *population, unemployment rate, migration in, migration out, economic data (revenues of companies, margins, tax), highly educated people, English speaking people at least on middle level, cultural facilities (cinemas, theaters), sport facilities, and so on.* in a table form. Many thematic map can be constructed based on this valuable database that is important for decision support in local government, local policy. As many data columns as many kind of thematic maps. This logic takes only one data column into account. In most cases this functionality is enough for the everyday requirements.

There can be such a situation where we are going to take more than one or all the data column into consideration for a serious analysis. In that case we are willing to make classes based on every data column. Unfortunately the traditional thematic mapping technique does not support this function because the algorithm works on only one data column.

What is the problem with the traditional thematic mapping? When we make a simple thematic map the algorithm picks only one data column up for the computation of classes, consequently other data columns are ignored in the definition of classes. This is a huge waste of data. If the problem that we are going to investigate is simple and the one-column-based thematic map represents every effect, this technique is acceptable. If the problem is multivariate, the understanding of it requires more than one data column, the traditional thematic mapping becomes unusable and unacceptable. Any market leader GIS software can not serve solutions for this case.

The technique that will be introduced in this paper is willing to outline a method that resolves the multivariate thematic mapping problem.

1.2 Clusters, segments, classes

The modern database technology produced huge databases, so it is a hard task to find something in them. The name of the technology which helps us to navigate in huge databases is the data mining. These methods support the segmentation of huge databases into smaller units such as classes, clusters, data groups. One of the most important tool in data mining is the clustering, that sometimes called classification or segmentation. These are not synonymes but very similar. The aim of the data mining is the interpretation of data. The interpretation often means

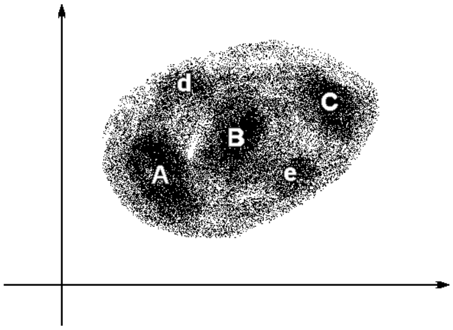

Figure 1.1: Classification by eyes (quick-look method). The A, B, C and the d, e groups can be recognized very well. It can be seen, that any point belonging to a group is nearer the center of its group then any other groups.

identification of data groups with memberships of data records. This procedure is the segmentation where we identify the membership of data records. If we know the physical meaning of the segments let we name it supervised classification. If do not know that meaning, but the data groups are made, let we name it unsupervised classification.

This technology is well known in the raster based GIS such as satellite image processing. Why do not we use clustering in vector based GIS in multivariate cases?

1.2.1 Definition of clusters

Before the definition of clusters look at the figure 1.1, where a point cloud can be seen based on the cross-plot technique, that figures two data columns in two dimensional space. It is simple to notice point-groups, that we can consider classes, at first look. If we investigate the point cloud we define three main groups (*A, B, C*) and two smaller ones (*d, e*). Look at the figure 1.1, where belonging points are near each other. Obviously this segmentation technique is not exact because the result is depends on the abilities of the person who look at it. In the other hand „quick-look" method works two dimensional cases only. If the dimension number is higher than two, the method becomes invalid.

If we are willing to define groups exactly, at first we should define similarity which will be the bases of clustering. Let us say, if two points are similar let they belong to the same cluster. The next question is when two points are similar? They are similar if they are near each other.

It must be emphasized that there are many different clustering algorithms. Our purpose is to introduce the possibilities of the application of clustering methods in vector based GIS. Look at some base concept on it.

1.2.2 Distance matrix

Let us define the distance between two data points, where u, v represent the certain points. The Euklidean-distance is the following:

$$d(u,v) = \left(\sum_{i=1}^{d} [u_i - v_i]^2 \right)^{1/2} .$$

Compute the distances between every data point. Let us denote the distance d_{ij}, between i-th and j-th points. Arrange distances into a matrix form, that is the distance matrix ($\underline{\mathbf{D}}$):

$$\underline{\mathbf{D}} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{pmatrix}$$

Near points are similar, so the distance matrix expresses the similarity. How to define clusters based on the similarity matrix? There are two main algorithm families for establishing clusters. The first algorithm group is the *partitioning procedures*, and the second group is the *hierarchic procedures*.

Partitioning procedures identify clusters with iterative approximation. The next pending point will be set up that group whose distance from the representative point (for instance weight-point) is the smallest one. In this way the cluster who includes that point, has to be recomputed with this new point, consequently the representative point will be changed a little. Every new point in a cluster makes little changes in the representative point of a certain cluster.

In the hierarchical procedures the data elements are arranged into trees where the data are is leaves, and inner points of the tree represent a cluster. We have two starting possibilities. The first one is that every point defines an individual cluster. Progressing the clustering process clusters have been unified depending on the conditions of the iteration.

The second starting possibility is that all data belongs to one cluster. If the clustering process is progressing the clusters are cut into smaller clusters depending on the conditions of the iteration.

1.3 Principal components

The cluster analysis takes all data into consideration while identifies clusters. If the dimension of the database is rather high, then the size of the distance matrix becomes seriously high, which can produce practical difficulties of the computation. In case of large dimensions practical difficulties make the computation impossible. There is only one way to solve clustering, to make dimensions decreased. The reduction of dimension number produces data waste. In this section an effective method will be described for decreasing dimension number which is the principal component analysis.

Let us have p observation vectors, and each vector has n data. Let $\mathbf{X^j}$ vectors be random variables and elements of them are realization of a random variable.

$\mathbf{x}^1$	$\mathbf{x}^2$	$\dots$	$\mathbf{x}^p$
x_1^1	x_1^2	$\dots$	x_1^p
x_2^1	x_2^2	$\dots$	x_2^p
$\vdots$			$\vdots$
x_n^1	x_n^2	$\dots$	x_n^p

In order remove physical dimensions from data let them be standardized

$$\tilde{x}_i^j = \frac{x_i^j - \overline{x}^j}{s^j}$$

where $\overline{x}^j$ is the mean of j-th vector elements and $\tilde{s}^j$ is the empirical scatter. In this way random variables became 0 mean and 1 empirical scatter (figure 1.2).

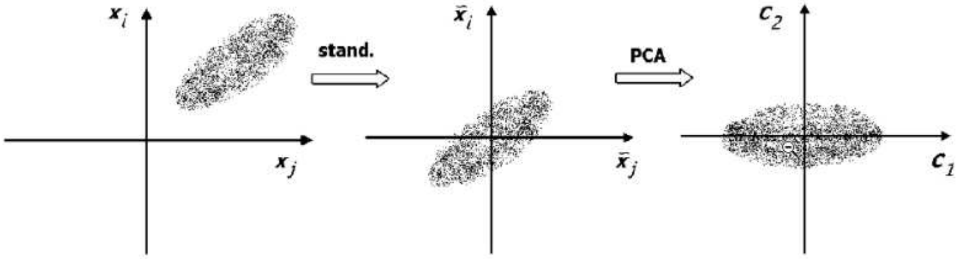

Figure 1.2: Geometrical meanings of a the standardization and the principal component transformation. The standardization makes dataset 0 mean and 1 empirical standard deviation. Principal component transformation moves the point cloud to the origo and rotate it to the proper direction.

Compute the correlation matrix ($\underline{\mathbf{R}}$) of the dataset, where

$$r_{xy} = \frac{M[(x - M(x))(y - M(y))]}{D(x)D(y)}$$

$$\underline{\mathbf{R}} = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{pmatrix}$$

Let us compute the eigenvalues and the eigenvectors of the correlation matrix, i. e. solve the following equation:

$$\underline{\mathbf{R}}\mathbf{v} = \lambda\mathbf{v}$$

Denote $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_p|$ the eigenvalues and $\mathbf{v^1}, \mathbf{v^2}, \ldots \mathbf{v^p}$ the eigenvectors. Based on the eigenvectors and the standardized random variables the j-th principal component can be computed

$$C_i^j = \sum_p x_i^p v_p^j$$

where $i = 1, 2, \ldots p$ and $j = 1, 2, \ldots p$.
Principal components have some important properties, such as

- principal components make orthogonal system, i.e. uncorrelated. Their correlation matrix is diagonal

$$\underline{\mathbf{R}} = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{pmatrix}$$

- eigenvectors are normalized, i.e. theirs scalar product is $< v^i \cdot v^j > \delta_{ij}$, where δ is the Kronecker-δ.
- The sum of eigenvalues is equal to the numbers of observation vectors.
- The variance content of the standardized variables and the principal components are the same, i.e.

$$\sum_{j=1}^{p} \lambda_j = \sum_{i=1}^{p} \tilde{s}_i^2 = \sum_{k=1}^{p} s_k^2 = p$$

Finally we can establish, that the principal component analysis rearranged the variances and extracted it to the first principal component. The geometrical meaning of the principal component analysis can be seen on the figure 1.2.

If the first principal component includes the largest common part of variances, it can substitute whole dataset in the certain computation. Based on this method we neglect some part of variances, but the waste is optimal, i.e. not more then it has to be. This technique does not work if the original dataset is uncorrelated. In this case we need to think about our data model and reconstruct it.

The introduced method is based on the random variable approach, but it can be defined on algebraic bases also.

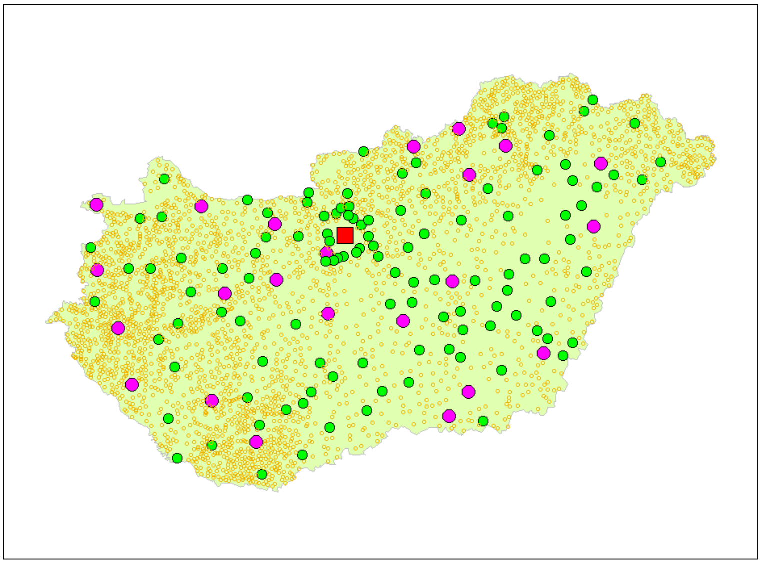

Figure 1.3: This thematic map shows the standard population of the settlements.

1.4 Thematic map as a result of clustering

The previous logic improved the importance of cluster analysis, and in case of large number of dimensions, the practical benefits of the application of principal component analysis. Thematic mapping is an everyday task in the vector based GIS, if we need a segmented database into classes in order to interpret data easily. Market leader GIS software packages unfortunately do not support the direct cluster analysis despite of raster based satellite image processing systems. In order not to give up the benefits of taking every data column into account, if we are willing to make dataset to be segmented, we should use the principal component analysis in case of traditional thematic mapping.

Based on the first principal component we can construct a thematic map on it. Compare the traditional method (figure 1.3 and 1.4) to the first principal component based one (figure 1.5).

Sometimes we need to understand the physical meaning of principal components, that inform us the background of dataset. Regarding the content of the original dataset it can be identified well. Sometimes it can not be done, because physical effects behind the phenomena can not be interpreted. If we are going to make thematic maps on the dataset we should understand the meaning of the first principal component.

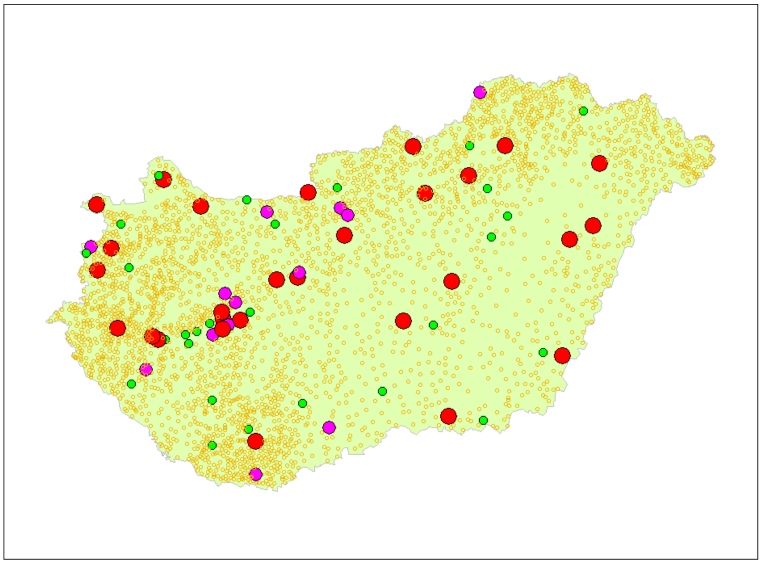

Figure 1.4: This thematic map shows an important touristic information: the sum of nights in hotels at the certain settlement.

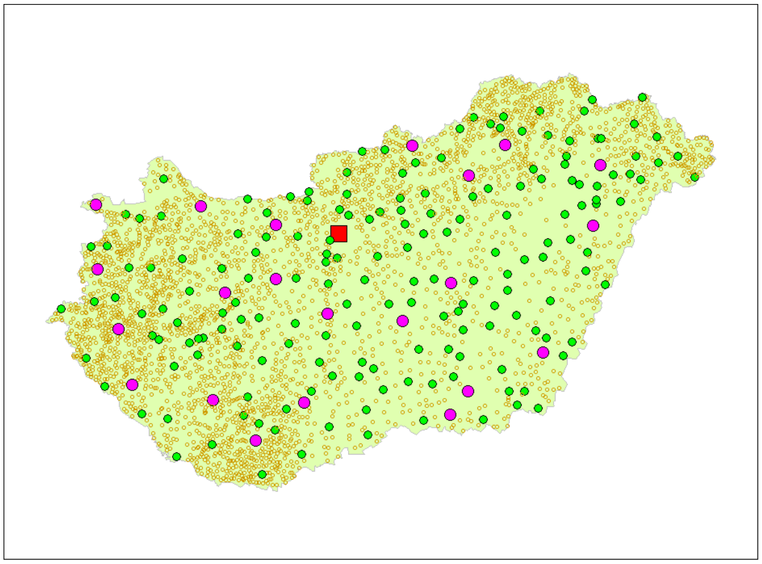

Figure 1.5: A thematic map that is based on the first principal component. The first principal component of settlements is coming from many statictical data of them. The result of the classification is the same than the status of settlements (figure 1.6)

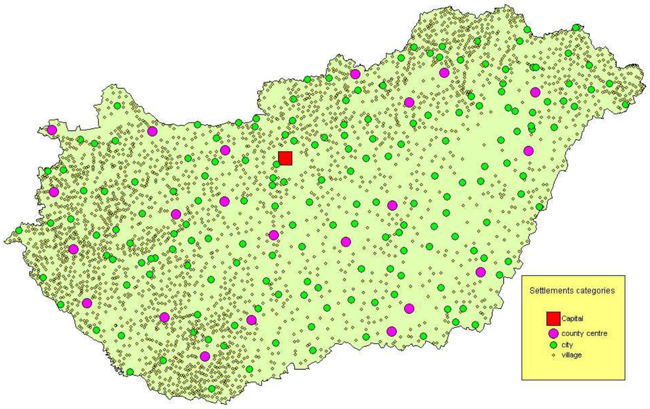

Figure 1.6: A thematic map is based on the status of settlements (capital, center city of a county, city, village, others). The result of the classification is the same as on the first principal component (figure 1.5)

1.5 Conclusion

The traditional thematic mapping is an excellent tool if we are willing to express graphically the spatial distribution a certain data column. If our classification analysis requires more than one data columns taking into consideration we should use cluster analysis or principal component analysis if we have traditional thematic mapping functionalities only. The resulted classes are more reliable than the single data column based technique.

Bibliography

[1] Ivanyi A. (editor) „Informatikai algoritmusok 1-2.”, ELTE Eotvos Kiado, 2005

[2] A. Stein – F. Meer – B. Gorte (editor): Spatial Statistics for Remote Sensing, Kluwer Academic Publishers, 1999

[3] I. Elek: „Fast Porosity Estimation by Principal Component Analysis”, GEOBYTE, june 1990, Tulsa, Oklahoma

[4] I. Elek: „Some Applications of Principal Component Analysis: Well-to-Well Correlation, Zonation”, GEOBYTE, may 1988, Tulsa, Oklahoma

[5] J. F. Richards: „Remote sensing Digital image analysis”, Springer-Verlag, 1986, Australia

[6] Vincze I.: „Matematikai statisztika”, Tankonyvkiado, Budapest, 1980

[7] J. Davis: ”Statistics and Data Analysis in Geology”, John Wiley & Sons, Inc., 1973