Pattern-based approach to knowledge extraction from giga-cell geospatial raster datasets

J. Jasiewicz[1], P. Netzel[2], T. F. Stepinski[3]

[1]Adam Mickiewicz University, Dziegielowa 27, 60-680 Poznań
Email: jarekj@amu.edu.pl

[2]University of Wroclaw, Kosiby 6/8, 51-621 Wrocław, Poland
Email: pawel.netzel@uni.wroc.pl

[3]University of Cincinnati, Cincinnati, OH 45221, USA
Email: stepintz@uc.edu

1. Introduction

There is a keen interest in development of automated methods of knowledge discovery from large geospatial raster datasets. A geospatial raster is large either because it has an extensive geographical coverage or because it has a very fine resolution. Standard GIS methods of analysis are ill-suited for such datasets as they operate either at the level of individual raster cell or at the level of multi-cell "object" (Blaschke, 2010). Such approaches will provide little insight in application to large (say, having 10^9 cells) raster. This is because the size of the cell and the data spatial extent differ by too many orders of magnitude. To interpret and analyze giga-cell rasters we propose an addition to the present GIS paradigms – a spatial analysis of local patterns. A local pattern (a "scene") is a mosaic of raster cell values. Importance of spatial patterns is in their association with specific geographic notions. For example, in a land cover dataset, a specific pattern of land cover categories can be recognized as a "downtown area." At smaller scale, an analyst could be interested in an actual configuration and composition of land cover categories constituting downtown, but at the large scale it's sufficient and indeed necessary to interpret this location by a single label. Traditionally, pattern analysis has been a domain of computer vision (Datta et al., 2008), whereas spatial analysis has been a domain of GIS. We have integrated the two domains by developing a methodology that enables GIS processing of local patterns.

2. Pattern-based GIS analysis of rasters

2.1 Defining scenes

Scenes are extracted from a raster in three different manners (see Fig.1B). (a) From neighbourhoods around a given set of points. (b) From a set of irregular segments. (c) As a regular grid-of-scenes. A grid-of-scenes covers the extent of original raster but has a much larger cell size. Grid-of-scenes stores pattern information extracted from square scenes centered on its cells; scenes overlap if they are larger than grid-of-scenes cells. A scene is represented by a pattern signature which encapsulates its character. Signatures are stored in the grid-of-scenes. We use histograms of pattern features as signatures. Pattern features are pieces of information about a pattern; only categorical features can be histogramed so numerical features must be categorized before they can be used. No single set of features describes well all possible patterns. GeoPAT – the software implementation of our methodology – implements several different methods of feature extractions.

2.2 Comparing scenes

In order to conduct operations on scenes we define a "distance" between the two scenes to be a distance between histograms of their features. There is no a single, universally accepted measure of distance between two histograms. Moreover, distance function needs to be matched with features selected for scene signature. In GeoPAT we have implemented several distance functions that match well with types of features we use. For example, to quantify patterns of land cover we use two

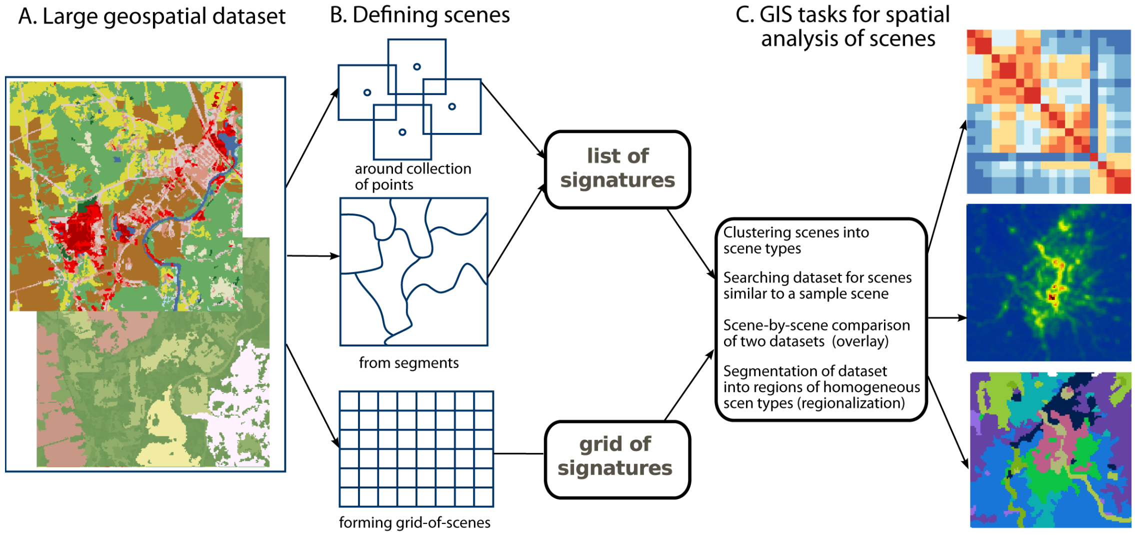
A. Large geospatial dataset
B. Defining scenes
C. GIS tasks for spatial analysis of scenes
around collection of points
from segments
forming grid-of-scenes
list of signatures
grid of signatures
Clustering scenes into scene types
Searching dataset for scenes similar to a sample scene
Scene-by-scene comparison of two datasets (overlay)
Segmentation of dataset into regions of homogeneous scen types (regionalization)

Figure 1: Application of pattern-based analysis to 30m/cell DEM covering the entire extent of the country of Poland.

features (land cover category and categorized size of clump to which a given cell belongs (Jasiewicz and Stepinski, 2013a) and we measure distance between histograms using Jensen-Shannon divergence (Lin, 1991). To quantify topographic patterns we use co-occurrence features (Haralick et al., 1973) and the Wave Hedges distance.

2.3 GIS processing of patterns

GIS tasks are performed on the grid-of-scenes with a scene signature serving as a new type of cell attribute and a query-by-pattern-similarity (QBPS) based on histogram distance replacing the SQL. Specifically, we have implemented four GIS tasks for spatial analysis of scenes (see Fig.1C). (a) Clustering of scene collection (for example, into landscape types or physiographic types). (b) Search of dataset for scenes similar to a given sample pattern with results visualized in terms of similarity map that reveals a geospatial context of the sample pattern. (c) Comparison between two co-registered grids-of-scenes which is equivalent to a standard GIS overlay function and can be used for assessing change in local patterns. (d) Segmentation of grid-of-scenes into sub-regions of uniform patterns. This task performs regionalization of a raster with respect to patterns of original variable. For example, if applied to a land cover dataset it will map its constituent landscape types, and, when applied to a topographic data, it will map its physiographic regions.

3. Application to topographic dataset of Poland

To demonstrate our methodology we have performed all four tasks described in section 2.2 using a 30m/cell DEM (21,696 × 24,692 cells) covering the territory of Poland. First, the DEM was categorized into 10 landform elements using the geomorphons method (Jasiewicz and Stepinski, 2013b). The landform elements map is shown in the center of Fig.2. We constructed a grid-of-cells with the resolution of 1.5 km (50 times larger than a resolution of the DEM) and extracted (overlapping) scenes having size of 15×15 km each. We used co-occurrence features to construct histograms and the Wave Hedges function to measure distances between histograms. Panel A shows schematically the results of the clustering task. A collection of scenes is clustered using hierarchical clustering method (shown here graphically in a form of a “heat map”). In this example the collection splits into three clusters interpreted as mountains, hills, and lowlands, respectively. Panel B shows the working of the search task. Three queries are shown each resulting in a creation of similarity map visualizing spatial extent of terrain similar to the corresponding query. Panel C shows a comparison between two rasters. In this case both rasters are categorizations of the original DEM but using different parameters, the result is a map showing regions where change in categorization parameters

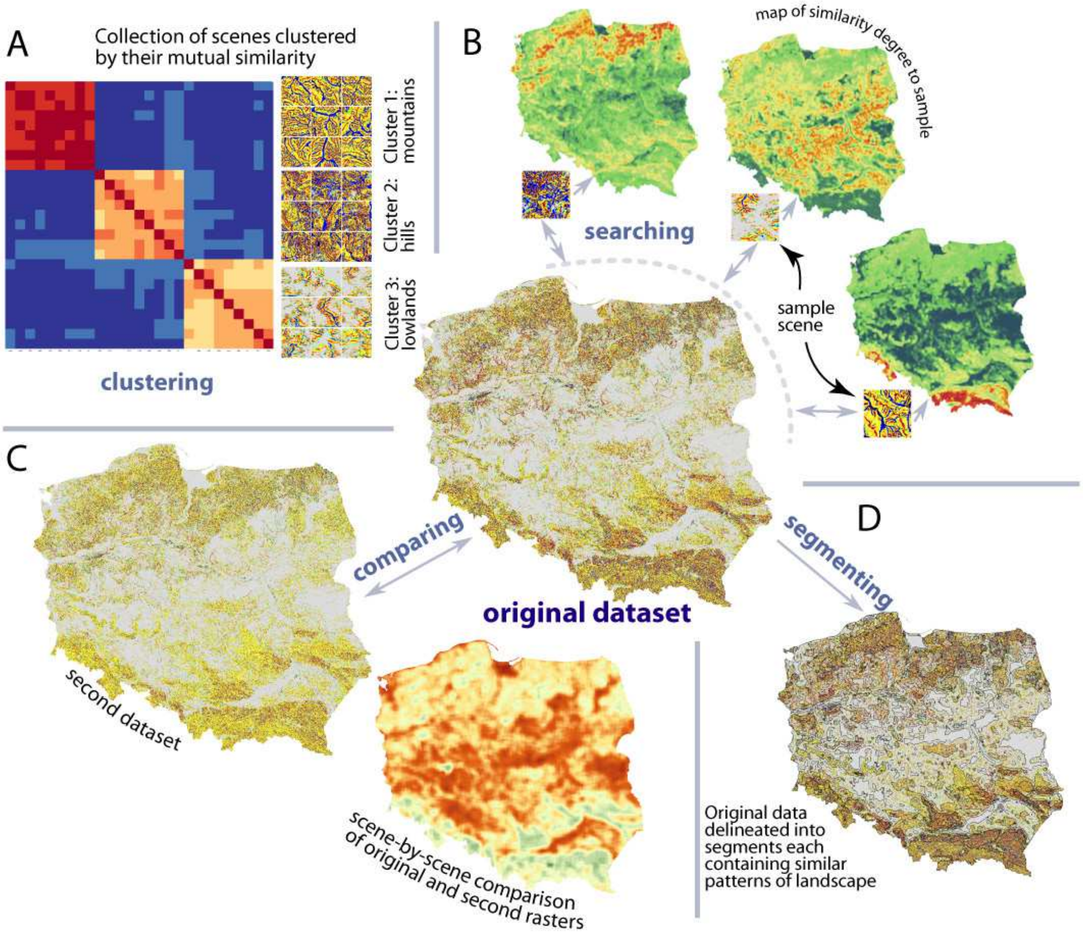
A
Collection of scenes clustered by their mutual similarity
Cluster 1: mountains
Cluster 2: hills
Cluster 3: lowlands
clustering
B
map of similarity degree to sample
searching
sample scene
C
comparing
original dataset
second dataset
scene-by-scene comparison of original and second rasters
D
segmenting
Original data delineated into segments each containing similar patterns of landscape

Figure 2: Application of pattern-based analysis to 30m/cell DEM covering the entire extent of the country of Poland.

results in significantly different interpretations of landscapes. Panel D shows results of the segmentation task. The territory of Poland is divided into irregular segments so that topographic pattern (landscape) in each segment is uniform. Classification of those segments yields a physiographic map of Poland (see Jasiewicz et al., 2014 for details). Such map could be thought of as the result of unsupervised machine learning. Alternatively, by selecting samples of major physiographic regions in Poland and running search task, a supervised learning-based physiographic map of Poland can be created as well.

3. Conclusions

We presented a methodology for knowledge extraction from giga-cell raster datasets. The methodology is based on spatial processing of local patterns formed by an original dataset variable. Computer visions concepts of pattern representation and similarity are used to extend several GIS tasks to work with grids of patterns. All steps necessary to apply our methodology in practice are contained in the toolbox GeoPAT which is freely available for download from http://sil.uc.edu/gitlist . This toolbox is a collection of newly written GRASS GIS modules. Integration with GRASS assures that very large datasets can be processed with ease and makes possible creation of scripts for computational pipelines that use other GRASS, as well as R modules. Such integration make complex tasks, such as, for example, supervised or unsupervised delineation of physiographic regions over large extents, relatively fast and easy. In addition to land cover and topography, our method can be applied to multiple other domains including crops, soils, climate, phenology, as well as socioeconomic data. It can also be applied to very large, high resolution images where multiple instance learning approach (Vatsavai, 2013) is currently used for the segmentation task.

Acknowledgements

The work was supported by NSC GRANT DEC-2012/07/B/ST6/01206, by the NSF under Grant BCS- 621 1147702, and by the University of Cincinnati Space Exploration Institute.

References

Blaschke, T., 2010, Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens*. 65(1), 2–16.

Datta, R., Joshi, D., Li, J., Wang, J. Z., Apr. 2008. Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Comput. Surv. 40 (2), 1–60.

Jasiewicz, J., Stepinski, T. F., 2013a. Example-Based Retrieval of Alike Land-Cover Scenes From NLCS2006 Database. *IEEE Geosci. Remote Sens. Lett.* 10 (1), 155–159.

Jasiewicz, J., Stepinski, T.F., 2013b. Geomorphons-a pattern recognition approach to classification and mapping of landforms. Geomorphology 182, 147–156.

Jasiewicz, J., Netzel, P., Stepinski, T.F. 2014. Landscape similarity, retrieval, and machine mapping of physiographic units. Geomorphology 221, 104–112.

Haralick, R. M., Shanmugam, K., Dinstein, I., Nov. 1973. Textural features for image classification. Syst. Man Cybern. IEEE Trans. 3 (6), 610–621.

Lin, J., 1991. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory 31(1), 145–151.

Vatsavai, R.R. 2013. Gaussian multiple instance learning approach for mapping the slums of the world using very high resolution imagery. In Proceedings of the 19th ACM SIGKDD, 1419-1426.