RETRIEVAL OF PATTERN-BASED INFORMATION FROM GIGA-CELLS CATEGORICAL RASTERS - CONCEPT AND NEW SOFTWARE

Jaroslaw Jasiewicz[1]*, *Pawel Netzel*[2], *Tomasz F. Stepinski*[3]†

[1]Adam Mickiewicz University, Institute of Geoecology and Geoinformation,
Dziegielowa 27, 60-680 Poznan
[2]University of Wroclaw, Department of Climatology and Atmosphere Protection,
Kosiby 6/8, 51-621 Wroclaw, Poland
[3]University of Cincinnati, Space Informatics Lab, Department of Geography,
Cincinnati, OH 45221-0131, USA

Index Terms— Information retrieval, giga-scale geocomputation, similarity analysis

1. INTRODUCTION

Rapid development of computer technology together with the growing availability of giga-scale data sources brings new possibilities to geo-spatial analysis [1, 2]. We define giga-scale datasets as those having size exceeding 10^9 cells, regardless of their physical scale. They may represent local regions at ultra-high resolution (of the order of centimeters) offered by LiDAR technology or global mosaics of satellite imagery or digital elevation models (DEMs) at medium resolution (of the order of 10-100 meters). Frequently, these giga-scale datasets are categorical rasters - products derived from processing of original data. Examples include land cover/land use (LCLU), landforms, vegetation, and urban maps. In such rasters important information is stored not only at the level of individual cells, but also, and maybe predominantly, at the level of patterns of the categories [3, 4]. Urban structures, plant habitats, geomorphological surfaces, and landscapes are examples of such patterns; they have collective functions and meaning and thus contain valuable information that cannot be inferred at the level of cell-based analysis.

Analyzing patterns, especially patterns in giga-scale datasets is not feasible by means of visual inspection but it is feasible by means of parsing the data by "intelligent" algorithms. In this paper we present a GeoPAT (Geospatial Pattern Analysis Toolbox) – a conceptual framework for retrieval of such pattern-based information and an implementation of this framework in a ready-to-use software. The framework is applicable to all categorical datasets and its use is illustrated using a raster containing categories of landform elements derived from a DEM.

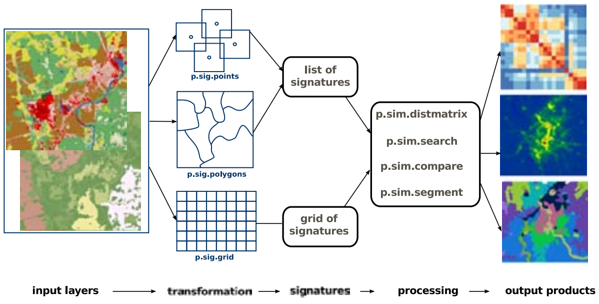

Fig. 1. Different data processing pipelines using GeoPAT toolbox. Input contains GRASS raster layers and output contains raster layers or text tables

2. THE CONCEPT

In our approach a basic areal unit of analysis is referred to as a "scene." We define a scene as a part of the dataset (Fig. 1) contained in a local region (square circle or irregular) of a given size. Thus scene is a local pattern of categories represented in a dataset. Scene size should be small in comparison to the extent of the entire dataset but it should contain a large number of individual raster cells to represent a meaningful pattern. The lattice of all scenes covers the same spatial extent as the original raster but the number of areal units (scenes) in such lattice is much smaller than the number of areal units (cells) in the original raster. Scene signature is defined as a nominal histogram of pattern features; different pattern features maybe selected for scene signature depending on the character of the data. Similarity between two patterns is calculated as a similarity between their signatures; different similarity measures may be selected for signatures built using different features.

Scene signatures and similarity measures are the two basic blocks of our framework. From these blocks we have designed and implemented software tools to address the four most frequent investigative needs: 1) **clustering**: similarity

[*] This work was founded by the Polish National Centre of Science under grant DEC-2012/07/B/ST6/01206

[†] This work was founded by NSF GRANT BCS-1147702.

978-1-4799-5775-0/14/$31.00 ©2014 IEEE

analysis between separate scenes; 2) **searching**: looking for particular scenes in entire spatial dataset; 3) **comparing**: spatial analysis of differences between co-registered datasets; 4) **segmentation**: segmenting the entire dataset into mutually exclusive and exhaustive regions, each grouping patterns that although unique in their details, are similar on a broader level.

The goal of clustering is to look for regularity (or lack of it) in the scenes constituting a given region or any other set of the scenes. Clustering tool calculates a similarity/distance matrix between all the scenes. Further analysis, such as, various forms of clustering or visualization can be performed using that matrix. Searching results in creation of new layer of information, having the same spatial extent and granularity as the lattice of scenes. Each cell of this layer contains a similarity value between a user-selected query scene and all scenes included in the dataset [5, 6]. Previously we have implemented searching as the GeoWeb app for query-and-retrieval of LULC pattern across the United States (LandEx-USA [7]) and as the GeoWeb app for query-and-retrieval of terrain landscapes across the country of Poland. These two apps are accessible from http://sil.uc.edu/. The goal of comparing is to calculate similarity between individual spatially co-registered scenes in two different co-registered lattices of scenes. The most obvious application of this functionality is for pattern-based change detection [8]. In pattern-based change detection the change occurs if the pattern changes; the change is very small if the patterns at two time steps are very similar even if raster categories have changed in a large number of cells. In addition to change detection, comparing can be utilized for assessment of two categorical products created with different parameters. Finally, regionalizing allows for delineation of regions containing scenes with similar patterns. It can be utilized for finding LULC landscape types [9] from land cover/land use datasets or physiographic provinces from landform datasets.

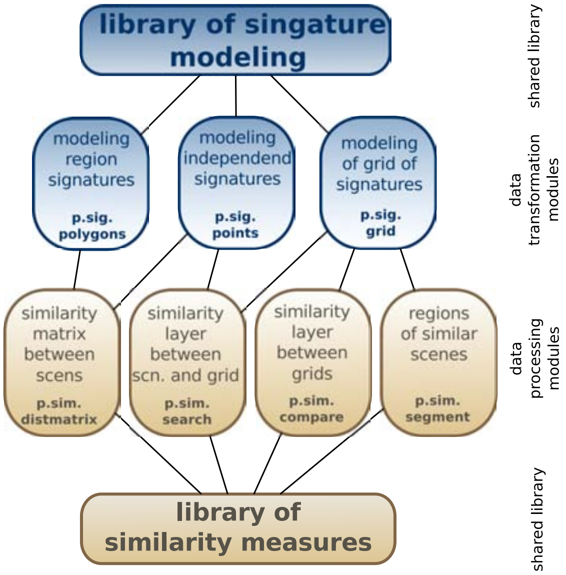
library of singature modeling
modeling region signatures
p.sig. polygons
modeling independend signatures
p.sig. points
modeling of grid of signatures
p.sig. grid
similarity matrix between scens
p.sim. distmatrix
similarity layer between scn. and grid
p.sim. search
similarity layer between grids
p.sim. compare
regions of similar scenes
p.sim. segment
library of similarity measures
shared library
data transformation modules
data processing modules
shared library

Fig. 2. A structure of similarity toolbox

3. THE SOFTWARE

The framework for pattern-based information retrieval from categorical rasters is implemented as *GeoPAT* Geospatial Pattern Analysis Toolbox (Fig. 1,2) including seven modules written in ANSII C for GRASS GIS environment. GRASS environment is used because it provides clear and mature APIs and is publicly available. The central part of our toolbox are the two libraries shared by all modules: histogramlib which contains routines for scene signature modeling and similarlib which calculates similarity between scenes. The histogramlib currently offers four routines for constructing histograms using four different pattern features: 1) *crossproduct* which uses features described in [10, 5, 7]; 2) *coocurence* where features are the pairs of categories [11, 12, 6]; 3) *decomposition*, loosely based on the concept described in [13] and [14]; 4) *multidimentional cdf*, designed for ordinal histograms. The similarlib library contains several similarity measures [15] including: Jensen-Shannon, Wave-Hedges, Róžiczka, Czekanowski, L1 and L2 Minkowski metrics or Chi-Square.

Three modules included in the toolbox are designed for data pre-processing (Fig. 1): **p.sig.points** is a tool for rapid modeling of example scenes defned by point coordinates, white **p.sig.polygons** is designed to work with scenes defined by irregular regions. Finally **p.sig.grid** is intended to work with large amount regullary arrange scenes. Toolbox works in parallel with two grid systems: $\mathcal{G}_i$ which contains input data and $\mathcal{G}_j$ which store results. The resolution and extend of the $\mathcal{G}_i$ is managed by GRASS GIS system [16] and describes the granulation of the input data. Structure of $\mathcal{G}_j$ inherits $\mathcal{G}_i$ extend, while its resolution is determined by the integer number which define the reduction of the original resolution. Size of the scene is controlled by another parameter and may exceed the resolution of the $\mathcal{G}_j$ grid, so adjacent scenes can overlap.

For other modules (Fig. 2)implement scenarios as defined in the 2. The **p.sim.distmatrix** produce distance matrix; **p.sim. search** realities all tasks included in the searching strategy; **p.sim.compare** is designed to study co-registered scenes while **p.sim.segment** allow to split large data sets into more or less uniform regions.

4. PERFORMANCE

To demonstrate performance of our software we run several examples on various small and big datasets using a computer with the Dual Xeon-8-core processors and 20GB of RAM. Average performance times are presented in the Table 1. We

task	input	output	time
single scene signature modeling using coocurence algorithm	single scene 300 × 300 cells with 10 categories	single histogram with 55 bins	0s 73ms
grid of histogram modeling using coocurence algorithm	data layer 24000 × 27000 cells with 10 categories	grid of histograms 800×900 cells with 55 bins	2h 26m
grid of histogram modeling using coocurence algorithm	data layer 84000 × 64000 cells with 10 categories	grid of histograms 1680 × 1280 cells with 55 bins	26h 39m
grid histogram modeling using crossproduct algorithm	data layer 164000 × 104000 cells, two input layers	grid of histograms 1640 × 1040 cells with 192 bins	3h 46m
clustering scenes	164 single histograms with 55 bins	a similarity matrix 164×164	1s 12ms
query-by-pattern similarity with Jensen-Shannon similarity measure	1 single histogram with 192 bins and 1640×1040 grid of histograms	one 1640 × 1040 similarity layer	6s 11ms
comparison of two co-registered grids of histograms using Wave-Hedges similarity measure	two 800 × 900 grid of histograms	one 800 × 900 similarity layer	0s 97ms
segmentation of single grid of histograms with Euclidean similarity measure	one 800 × 900 grid of histograms	one 800 × 900 layer of regions with hemogenous pattern	1s 28ms

Table 1. Comaprison of various tasks performed by similarity toolbox

also optimized **p.sig.grid** module through the application of parallel computing using OpenMP library. Significant differences between frequently executed tasks like scene modeling or various retrieval scenarios (orders of seconds) and rarely run data preparation (order of hours) as well as modular construction of the library allow to use our software in practical implementations [7] and as a framework for complex scientific analysis [6].

5. CONCLUSION AND FUTURE WORKS

The central idea of our concept is to perceive a raster containing a very large number of cells, each carrying a single value, as an another raster containing much smaller number of larger cells, each representing a local spatial pattern of original values. We call this new raster a grid-of-scenes. An important property of the grid-of-scenes is that it can be processed by a GIS system much like any other grid providing that its cells carry a new type of attribute – scene signature. GeoPAT extends a GIS system by providing modules and functions to build a grid-of-scenes form original raster, to calculate scene signature attribute, and to geoprocess a grid-of-scenes in a manner which is transparent to a GIS user. Standard grid processing operations, such as, search, overlay, and segmentation, are performed on the grid-of-scenes in a usual fashion even so, behind-the-scenes, GeoPAT uses a pattern-specific query system instead a traditional SQL system. GeoPAT makes possible spatial analysis at the higher level of abstraction (pattern vs. cell), thus we expect that by using it analysts would be able to address questions which would be difficult to even formulate without it.

GeoPAT is an actively developed software, presently at the very beginning of its development cycle. We expect that users will contribute to GeoPAT project by making core modules more stable and by adding to the shared library of functions. Roadmap for future development of GeoPAT is as follows:

Expansion of shared library. Existing shared library of pattern signatures and distance functions reflects our own hands-on experience which is restricted to land cover and topography data having resolutions of 30-90 m/cell. We plan to experiment with other types of data including ∼ 1m/cell image and LIDAR datasets, as well as census grids, soil grids, climate grids, etc. This will result in development of new signatures and distance functions which will be added to the library.

Improvements and expansions of core modules. At present search and segmentation of grid-of-scenes are restricted to single algorithms. We plan on expanding the module p.sim.search to compound queries featuring AND and OR logical connectors. With respect to segmentation we plan on expanding p.sim.segment (which uses a region-growing method) by adding different methods of segmentation.

User experience. At present GeoPAT is available as a research tool whose current development concentrated on computational efficiency rather than ease of use or ease of installation. If the idea of pattern-based GIS will gain momentum we will develop a "customer" version of GeoPAT that is easier to install and use.

6. REFERENCES

[1] P. Netzel and T.F. Stepinski, "Connected components labeling for giga-cell multi-categorical rasters," *Computers & Geoscience*, vol. 59, pp. 24–30, 2013.

[2] P. Guth, "The Giga Revolution in Geomorphometry: Gigabytes of RAM, Gigabyte-Sized Data Sets, and Gigabit Internet Access," in *Geomorphometry 2013*, Nanjing, China, 2013.

[3] N. Vasconcelos, "From Pixels to Semantic Spaces: Advances in Content-Based Image Retrieval," *Computer*, vol. 40, no. 7, pp. 20–26, July 2007.

[4] M. Datcu, K. Seidel, S. D'Elia, and P.G. Marchetti, "Knowledge-driven information mining in remote-sensing image archives.," *ESA Bulletin*, vol. 110, no. may, pp. 26–33, 2002.

[5] J. Jasiewicz and T.F. Stepinski, "Example-Based Retrieval of Alike Land-Cover Scenes From NLCS2006 Database," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 1, pp. 155–159, 2013.

[6] J. Jasiewicz, T.F. Stepinski, and P. Netzel, "Content-based landscape retrieval using geomorphons," in *Geomorphometry 2013*, Nanjing, China, 2013.

[7] T.F. Stepinski, P. Netzel, J. Jasiewicz, and J. Niesterowicz, "LandEx–A GeoWeb-based Tool for Exploration of Patterns in Raster Maps," in *GIScience 2012 7th International Conference on Geographic Information Science*, Ningchuan Xiao, Mei-Po Kwan and Hui Lin, Eds., Columbus, OH, 2012, p. 5.

[8] T.F. Stepinski and P. Netzel, "Pattern-based assesment of 2001/2006 Land cover change over the entire United States," in *This volume*. This volume, IEEE.

[9] J. Niesterowicz and T.F. Stepinski, "Regionalization of multi-categorical landscapes using machine vision methods," *Applied Geography*, vol. 45, pp. 250–258, 2013.

[10] J.w Jasiewicz and T.F. Stepinski, "Query and retrieval of land cover patterns," in *IEEE International Geoscience and Remote Sensing Symposium*, 2012.

[11] Robert M.R.M. Haralick, K. Shanmugam, and Its'Hak Dinstein, "Textural features for image classification," *Syst. Man Cybern. IEEE Trans.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[12] M.J. Barnsley and S.L. Barr, "Inferring urban land use from satellite sensor images using kernel-based spatial reclassification," *Photogrammetric engineering and remote sensing*, vol. 62, no. 8, pp. 949–958, 1996.

[13] T.K. Remmel and F. Csillag, "Mutual information spectra for comparing categorical maps," *International Journal of Remote Sensing*, vol. 27, no. 7, pp. 1425–1452, 2006.

[14] R. Vetro, W. Ding, and D.A. Simovici, "Mining for high complexity regions using entropy and box counting dimension quad-trees," in *Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on*. 2010, pp. 168–173, IEEE.

[15] S.H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Appled Sciences*, vol. 1, no. 4, pp. 300–307, 2007.

[16] GRASS Development Team, *Geographic Resources Analysis Support System (GRASS GIS) Software*, Open Source Geospatial Foundation, USA, 2012.