LandEx - A GeoWeb tool for query and retrieval of spatial patterns in land cover datasets

Tomasz F. Stepinski, Pawel Netzel, and Jaroslaw Jasiewicz

***Abstract*—The vast amount of data collected by satellites via remote sensing is a valuable resource, however, it lacks machine search capabilities. In particular, large land cover datasets, such as the 30 m/cell NLCD 2006 covering the entire conterminous United States, are rarely analyzed as a whole due to the lack of tools beyond the basic statistics and SQL queries. Consequently, the NLCD is underutilized relative to its potential. We address this issue by introducing LandEx - a GeoWeb application for real time, content-based exploration and mining of land cover patterns in large datasets. By combining the functionality of online computerized maps with the power of the pattern recognition algorithm, LandEx provides an easy to use visual search engine for the entire extent of the NLCD at its full resolution. The user selects a pattern of interest (a query) and the tool produces a similarity map indicating the spatial distribution of locations having patterns of land cover similar to that in the query. Pattern-based query and retrieval addresses the issue of structural similarity between landscapes. The core of the method is the similarity function between two patterns which is based on 2D land cover class/clump size histograms and the Jensen-Shannon divergence. The search relies on exhaustive evaluation using an overlapping sliding window approach. LandEx is implemented using Free Open Source Software (FOSS) software and adheres to the Open Geospatial Consortium (OGC) standards. The wait time for an answer to a query is only several seconds due to the high level of system optimization. The methodology and implementation of LandEx are described in detail and illustrative examples of its application to different domains, including agriculture, forestry, and urbanization are given. LandEx is available at http://sil.uc.edu/landex/.**

***Index Terms*—pattern-based similarity, computerized maps, GeoWeb services, land cover datasets.**

I. Introduction

DURING the last decades, advances in remote sensing have made it possible to collect vast amounts of geospatial data. This deluge of data presents difficult challenges in terms of storage and distribution mechanisms. Although challenges in storage of big data have been overcome, effective distribution of this data to end users remains problematic. Large archives of remotely sensed data have been created (for example: The National Map Viewer (http://nationalmap.gov/viewer.html) or GeoBrain (http://geobrain.laits.gmu.edu/)), but intelligent methods of data retrieval are lacking. The most common way of accessing the content of large archives is to query them by geographical coordinates, time of acquisition, and sensor type. A consequence of such metadata-based approaches is that data retrieval from archives is restricted only to sites for which prior knowledge of their relevance exists. This severely limits the degree of utilization of existing archives; it is estimated that less than 5% [1] of all data available in remote sensing archives are actually used.

The solution is to develop a system capable of executing a content-based query that interprets the actual content of the data and returns instances that match the desired content. Such systems have been extensively studied (see reviews by [2], [3], [4]) in the context of natural image retrieval where they are referred to as Content-Based Image Retrieval (CBIR) systems. An image retrieval system contains a database with a large number of unlabeled images and an algorithm that retrieves images from the database on the basis of their intrinsic similarity to a query image entered by the user. The quality of the results varies depending on the content of the query image. Applications of CBIR to remotely sensed images in the geospatial domain have been also studied [5], [6], [7], [8]. In particular, a method for query and retrieval of satellite images was presented in [9]. To the best of our knowledge, the research on application of CBIR in geospatial domain is still in its early stages and no CBIR system for querying geospatial images is available in the public domain.

In this paper we describe the GeoWeb implementation of LandEx (Landscape Explorer) - a content-based map retrieval (CBMR) system. LandEx bears an overall design resemblance to CBIR systems, but it works on a different type of dataset and yields a different type of output. First, LandEx is designed to query categorical rasters (hereafter referred to as "maps") rather than images. The implementation of LandEx presented here utilizes the National Land Cover Dataset 2006 (NLCD 2006) [10]. Second, LandEx does not work with a database of many separate maps, instead, it creates its own database by subdividing the NLCD 2006 into a large number of (overlapping) tiles. It performs a search for spatial patterns of land cover classes using the query-by-example principle; a user selects a small tile from the NLCD 2006 as a query and the system calculates similarities between the query and all the other tiles in the NLCD 2006. Finally, the output of LandEx is not a short list of best-matching tiles, but rather a similarity map (of the same geographical extent as the NLCD 2006) showing a degree of similarity between the query and every other tile in the database. Such presentation of search results is more appropriate for spatial datasets where geographical context matters. The early version of our CBMR system was described in [11]. The system has since been

T.F. Stepinski is with the Space Informatics Lab., University of Cincinnati, Cincinnati, OH 45221-0131, USA, e-mail: stepintz@uc.edu

P. Netzel is with the Department of Climatology and Atmospheric Protection, University of Wroclaw, Kosiby 6/8, 51-621, Wroclaw, Poland.

Jaroslaw Jasiewicz is with the Geoecology and Geoinformation Institute, Adam Mickiewicz University, Dziegielowa 27, 60-680 Poznan, Poland

Manuscript received xxx xx, 2013; revised xxx xx, 2013.

further developed resulting in $\sim 10^3$ increase in search speed. This makes possible its implementation as a GeoWeb service. The GeoWeb implementation of our CBMR system is called LandEx and is freely available at http://sil.uc.edu/landex/; it returns query results in real time.

The ability to query a very large NLCD 2006 database in real time for specific patterns of land cover facilitates data-mining-like exploration of this database. LandEx makes it possible to pose questions about the spatial distribution of land cover that are impossible to even formulate without it. It makes utilization of the entire NLCD 2006 practical. The purpose of this paper is to describe LandEx and to demonstrate its abilities. The paper is organized as follows. Section II summarizes our approach to designing a CBMR system. Section III describes the implementation of LandEx. In Section IV we demonstrate potential uses of LandEx. Discussion and future work directions are given in Section V.

II. LANDEX METHODOLOGY

In LandEx a tile $\mathcal{A}$ is defined as a small subset of the entire NLCD 2006. For convenience we use square-shaped tiles with the size $n \times n$ cells. Each cell is labeled by one of $K = 16$ nominal labels $c_1, \ldots, c_K$ corresponding to the land cover/land use (LCLU) classes in the NLCD 2006. A query $\mathcal{Q}$ is a particular tile of interest selected by a user. There are two major components of LandEx methodology: the algorithm for calculating the similarity between $\mathcal{Q}$ and $\mathcal{A}$ and the execution of the query over the entire extent of NLCD 2006.

A. Similarity between two maps

A method of calculating an appropriate similarity value between two different maps is at the core of our methodology. Existing research on map comparison [12], [13], [14], [15], [16] pertains to the detection of temporal changes, to comparison between different mapping methodologies, to validation of LCLU models, and to assessment of map accuracy. Existing measures assign a high value of similarity to a pair of raster maps if the two maps show the same scene mapped from the same perspective with the only difference being somewhat different assignment of categories to corresponding cells. Note that such measures will assign a small similarity to the pair of identical maps if one is rotated 90 degrees with respect to the other. This is because even though the patterns of the two maps are identical, the corresponding cells will have different class assignments. Thus, existing methods are not relevant to the task of querying for similar spatial patterns where the two maps are expected to be assigned a high similarity value exclusively on the basis of an overall style or motif of spatial pattern without regard to relative rotation, translation, or some small degree of pattern deformation. Pattern-oriented similarity measures [17], [18], [19], [20] were considered in the context of landscape ecology using sets of landscape indices. The set of landscape indices [21], [22] was used in [20] to quantify spatial patterns in a map. However, it is not clear how to define a similarity function on the basis of a set of landscape indices. In [20] a set of 28 indices was calculated for each map in a collection of 182 LCLU maps and used to manually cluster the maps into 8 classes without defining an algorithmic measure of similarity.

In LandEx we use an original method [11] of calculating similarity between two maps. Our method is based on concepts developed in the context of CBIR domain. Like most CBIR methods our method has two components, pattern signature and pattern similarity. Pattern signature is a compact mathematical description of a pattern and pattern similarity is a function that assigns a numerical value of similarity between any two patterns (maps) on the basis of their respective signatures.

For pattern signature we use a class/clump-size histogram constructed from all cells in the map. Class (in the form of color) histograms are widely used in CBIR (for a review see [3], [23]) because of their rotational and translational invariance. One advantage of using class histograms in the context of maps is the small number of classes (16 for the NLCD 2006 versus 2^{24} colors for a typical natural image). Thus, a class histogram of a map can be constructed without using quantization. On the other hand, a class histogram accounts only for the overall bulk composition of classes in the map but not for their spatial pattern. In order to incorporate some spatial information into our signature we segment the map into clumps (four-connected region of a single class) using a standard connected components algorithm [24] and calculate the size of each clump in terms of the number of individual cells within it. Clump sizes are numerical data that upon quantization constitute the second component of our class/clump-size histogram. We quantize clump sizes by assigning them to bins with ranges based on powers of two (i.e. 1-2, 2-4, 4-8 etc). The number of bins, L, is determined by the size of the map (tile). We have experimented with adding shape of clump as an additional pattern descriptor. Using such reacher pattern signature did not enhance performance of a query while significantly lengthening its time of execution thus shape was not incorporated into a signature. Each cell inherits its clump-size class from the clump to which it belongs. The resulting 2D histogram remains invariant to rotation and translation. Fig.1 shows an example of constructing a class/clump-size histogram $A(\mathcal{A})$ from a tile $\mathcal{A}$. Fig.1A shows a tile $\mathcal{A}$ with the size $n = 500$ and $K = 16$ classes. Fig.1B shows the result of segmenting $\mathcal{A}$ into 1290 clumps (shown by assigning a random color to each clump). Fig.1C shows $A(\mathcal{A})$ with 16 land cover classes arranged along the x-axis and 14 clump-size classes arranged along the y-axis. The z-axis indicates a fraction of all the tile's cells that belong to a given class/clump-size bin. Because histogram $A(\mathcal{A})$ is normalized to unity it can be thought of as a probability density function (pdf) of a random variable X =(land cover class, clump-size class). Fig.1C indicates that the most likely outcome of variable X in tile $A(\mathcal{A})$ is "deciduous forest" in clumps having size class 12 and 9, and "developed open space" in clumps having size class 10 and 8. The computational cost of calculating histograms is dominated by the cost of segmenting the tile into clumps and varies depending on the complexity of a pattern. However, this cost is not a major issue because in LandEx calculation of histograms is performed off-line, they are pre-calculated and don't need to be evaluated during a query.

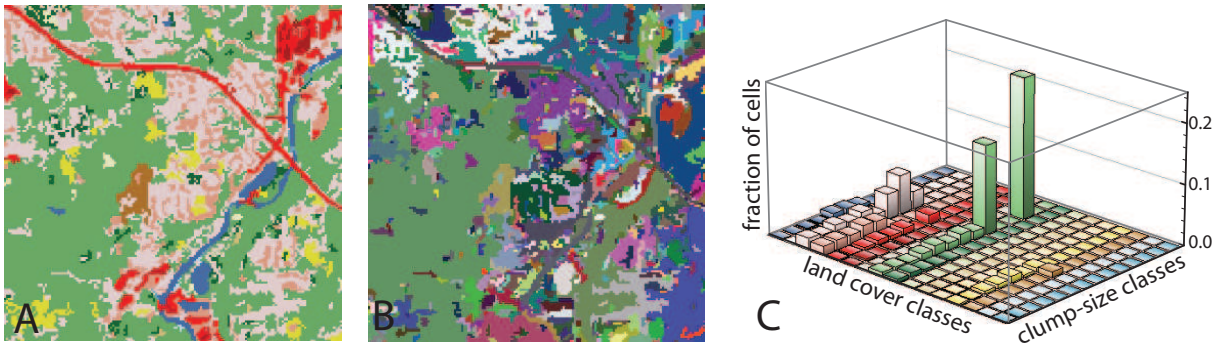

Fig. 1. Constructing a class/clump-size histogram. (A) An $n = 500$ (15km $\times$ 15km) tile of NLCD 2006; different colors indicate different land cover classes (see Fig.5 for a legend). (B) A tile segmented into clumps; random colors indicate individual clumps. (C) Class/size-clump histogram of the tile.

Calculation of similarities between two histograms is an on-line operation, each query requires $\sim 10^6$ similarity evaluations. In order to provide a real time response to a query, the similarity calculation needs to be very efficient. For this reason we consider our histograms to be of the nominal type where each bin is independent from other bins and the similarity of the non-overlapping parts of the two histograms do not need to be taken into consideration. There is a large selection of possible similarity functions, for a comprehensive survey see [25]. In LandEx we use the Jensen-Shannon divergence [26] to calculate similarity between two histograms. We have chosen the Jensen-Shannon divergence because of its robustness and good performance in side-by-side comparison with other measures [27]. In this context "divergence" is synonymous with dissimilarity or distance - a quantitative degree of how far apart the two histograms are. For two histograms A and B the Jensen-Shannon divergence (JSD) measures the deviation between the Shannon entropy [28] of the mixture of the two histograms $(A + B)/2$ and the mean of their individual entropies, and is given by

$$\mathrm{JSD}(A, B) = H\left(\frac{A+B}{2}\right) - \frac{1}{2}\left[H(A) + H(B)\right] \quad (1)$$

where $H(A)$ indicates a value of the Shannon entropy of the histogram A

$$H(A) = -\sum_{i=1}^{K}\sum_{j=1}^{L} A_{i,j} \log_2 A_{i,j} \quad (2)$$

where $A_{i,j}$ is the fraction of cells belonging to class i and clump-size j. JSD is always defined, symmetric, bounded by 0 and 1, and equal to 0 only if $A = B$. Note that JSD can be interpreted as the mutual information between variable X having distribution $(A+B)/2$ and a binary indicator variable Z where $Z = 1$ if X is from $\mathcal{A}$ (and has distribution A) and $Z = 0$ if X is from $\mathcal{B}$ (and has distribution B). Mutual information gives an average reduction in unpredictability (entropy) of X if the map is set.

The value of $H(A)$ reflects the distributional character of histogram A. A large value of $H(A)$ indicates A is evenly spread between the bins, whereas a small value of $H(A)$ indicates A is concentrated in just few bins. JSD measures (in a single number) the difference between the distributional characteristics of A and B. Note that if the two maps, $\mathcal{A}$ and $\mathcal{B}$, have similar histograms, $A(\mathcal{A})$ and $B(\mathcal{B})$, the histogram of their mixture, $(A + B)/2$, is similar to each of the two individual histograms and the value of JSD is small. If the two maps have dissimilar histograms, the histogram of the mixture is more spread than each of the two original histograms and the value of JSD is large. A maximum difference, JSD=1, is assigned for two histograms where each is having only a singe but different bin (two maps each having a single but different land cover class).

Taking advantage of the fact that JSD is bounded by 0 and 1 we define a similarity between A and B (and consequently a similarity between $\mathcal{A}$ and $\mathcal{B}$) as

$$\mathrm{JSS}(A, B) = 100\left[1 - \mathrm{JSD}(A, B)\right] \quad (3)$$

where JSS stands for Jensen-Shannon Similarity. The factor 100 is added so the similarity is expressed in terms of "percentage." Fig.2 shows an example illustrating how the values of JSS translate into visual similarity as perceived by a human analyst. The query tile is shown together with 15 other tiles having a similarity to the query in the range between 90% and 50%. We have arranged the retrieved tiles into three series in such a fashion as to show the perceived "correlation" between the values of JSS and the visual changes in the pattern. It is clear that as the value of JSS decreases from 100% the variation of patterns assigned the same level of similarity to the query increases. Interestingly, our system provides an example of the "Anna Karenina Principle" (AKP) [29]. According to the AKP, for something to succeed several key aspects or conditions must be fulfilled. Failure in any one of these aspects leads to failure of the undertaking. To match the pattern of a tile to a query many different features of the pattern must match. As tiles fail to match features of the query pattern they diverge from a perfect match in different ways. By performing hundreds of queries on NLCD 2006 we have determined empirically that for a retrieved tile to be visually perceived as highly similar to the query the value of JSS needs to be 90% or more.

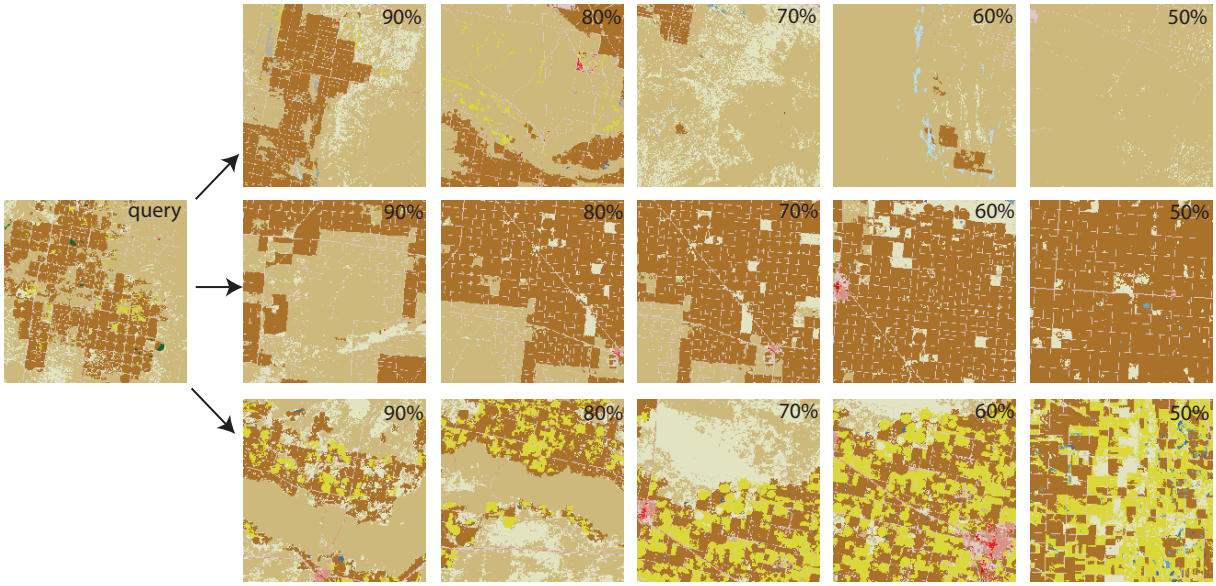

Fig. 2. Visual comparison of a query tile with fifteen local tiles having similarity to the query values between 90% and 50%. Local tiles are organized in three series in order to help in visualization between similarity value and visual divergence. See Fig.5 for a land cover legend.

B. Query Execution

For enabling a query over the entire NLCD 2006 we utilize an overlapping sliding window approach. A square grid with the resolution of k raster cells is superimposed over the entire spatial extent of NLCD 2006. This grid forms a basis for a similarity map resulting from the query. The query is executed by means of exhaustive evaluation - the value of JSS is calculated between the query tile and all the local tiles assigned to a similarity grid. Because our method compares maps using histograms, maps of different shapes and sizes can, in principle, be compared. However, in order to compare patterns on the same spatial scale, LandEx compares maps having square shapes and the same sizes. The size of the map tile is $N \times k$ to allow for overlap of neighboring local tiles. Thus, for example, for $k = 100$ raster cells and $N = 5$, all local tiles (and the query tile) have dimensions of 500×500 cells but the centers of neighboring tiles are only 100 cells apart allowing for ample overlapping.

LandEx is envisioned as a tool for rapid exploration of the entire NLCD 2006, so the queries need to be evaluated in real time. To make this possible we pre-calculate histograms for all local tiles off-line as they do not depend on the selection of a query. The on-line evaluation is restricted to calculating a single histogram of the selected query and to calculating the values of JSS(A^l,Q) between a histogram encapsulating the query $\mathcal{Q}$ and all the histograms A^l, $l = \{1, \ldots, M\}$ encapsulating local tiles, where M is the number of local tiles $\mathcal{A}^l$. The result of the calculation is a similarity grid having values in the range (0%,100%) indicating the degree of similarity between a local pattern and the query. The GeoWeb implementation of LandEx offers search capabilities on two different spatial scales, the coarser scale characterized by $k = 100$ and $N = 5$, and the finer scale characterized by $k = 50$ and $N = 3$. This limitation stems from the amount of computer memory needed to store the pre-calculated histograms. Note that the off-line version of LandEx has no spatial scale limitations, the query can be executed for any combination of parameters k and N, but the evaluation time is significantly longer due to a need to calculate histograms.

The coarser spatial scale search corresponds to comparing patterns of land cover over regions having a size of 15km $\times$ 15km with an overlap of 12km. These parameters have been chosen empirically; they offer, in our judgment, the optimal choice for US-wide similarity map. The resultant similarity raster has dimensions 1612 $\times$ 1045 and requires 1,684,540 evaluations of JSS for its completion. The finer spatial scale search corresponds to comparing patterns of land cover over regions having size 4.5km $\times$ 4.5km with an overlap of 3km. The resultant similarity raster has dimensions 3224 $\times$ 2090 and requires 6,738,160 evaluations of JSS for its completion. In practice both searches typically take only a few seconds to complete; the finer search does not take four times longer than the coarser search because the evaluation of JSS values is only one of many steps necessary to deliver a similarity map to the user. The wait time for return of a query may increase with increasing load on the server.

III. LandEx implementation

LandEx is a modern internet application running in a web browser. It is supported by most modern web browsers including Chrome, Firefox, Internet Explorer, and Safari. The computational part of LandEx is located on a dedicated server, whereas its interface (http://sil.uc.edu/landex) is provided via a web browser. Communication between the two parts is provided by the HTTP protocol. In other words, Landex is a computerized map application with all functionalities available through an active web page (like in, for example, Google Maps). As with most geospatial web portals, LandEx adheres to standards developed and published by the Open Geospatial

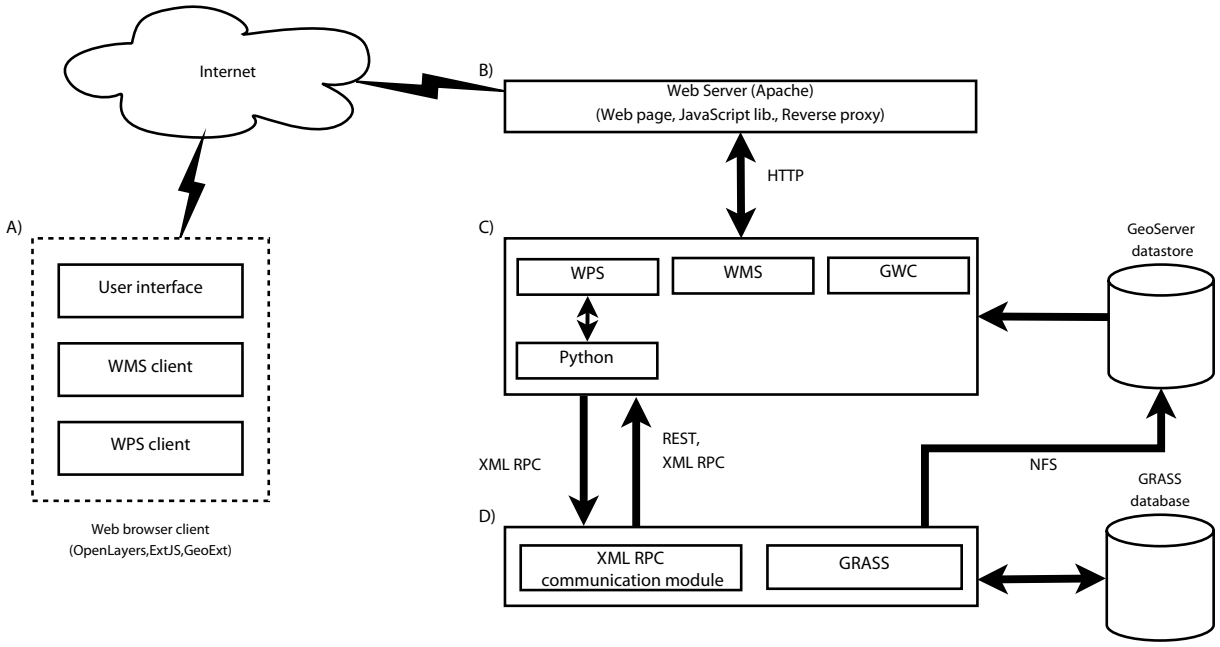

Fig. 3. Software architecture of LandEx. See main text for description.

Consortium or OGC (http://www.opengeospatial.org/). It uses two of the OGC standards: WMS (Web Map Service) and WPS (Web Processing Service). By using these standards we ensure that LandEx can be accessed not only through our own internet browser interface but also by third party software packages compatible with WMS and WPS protocols. This provides extra flexibility for utilizing the core functionality of LandEx.

LandEx is running on a server with the Linux system and all of its components are Free Open Source Software (FOSS) software. The architecture of LandEx is shown in Fig. 3; its main components are:

1) thin client in web browser environment;
2) web server;
3) map server which provides OGC services;
4) calculation engine based on the GRASS system.

LandEx's browser interface is based on JavaScript libraries: ExtJS with GeoExt and OpenLayers. Through this interface a user can access all functionalities (summarized in Fig. 4) expected from computerized map application. The server side of LandEx contains three components: the web server, the map server (working also as a map cache), and the calculation engine. Apache Web Server (http://httpd.apache.org/) is used as the web server component. It publishes the LandEx web page with user interface and provides the necessary JavaScript libraries. It is also utilized as a firewall so only selected OGC services are exposed to the extranet while the administration elements of the map server are restricted to the intranet.

GeoServer (http://geoserver.org) – an OpenSource OGC compliant software – is used as the map server. One advantage of using GeoServer is that it has built-in GeoWebCache (http://geowebcache.org/) application that provides pre-rendered map tiles. This assures highly efficient handling of map requests with low system load. All base maps (administrative boundaries, land cover, and shaded relief) are pre-rendered and cached using the GeoWebCache. We utilize two GeoServer plugins (WPS service and Python scripting) which are used to control the calculation engine.

LandEx calculation engine is based on the GRASS system (http://grass.osgeo.org/; see also book by [30]) and the XML-RPC server. Communication between the calculation engine and the map server uses the XML RPC protocol. The XML RPC-based communication module, custom built for LandEx, distributes requests from the map server to GRASS modules and controls the flow of results from the calculation engine to the map server. The most important function of the calculation engine is the computation of a similarity map for a given query. The processing flow is as follows: In the first step, a user selects the size and the position of a query region utilizing LandEx web interface. This information is sent to the map server which checks the request and sends it to the calculation engine. The calculation engine performs the similarity calculation and converts the results to the GeoTIFF format. Next, the GeoTIFF file is transfered to the GeoServer datastore where it is registered as a new layer. The map server returns the new layer to the web client which adds it into the layers tree and makes it visible to the user. The new similarity layer is available for the user for 20 minutes before being deleted to prevent the piling up of a large number of layers in the server memory, but the user is given the option to save the similarity layer as a GeoTIFF file.

The real time functionality of LandEx is due to optimization of system performance. Histograms data are pre-calculated and stored in separate files, so they don't need to be calculated during each query request. Calculation procedures use OpenMP library for the parallelization of the work. Network File System (NFS) protocol is used to achieve fast transfer of large files. As a result it takes only several seconds for LandEx to return a

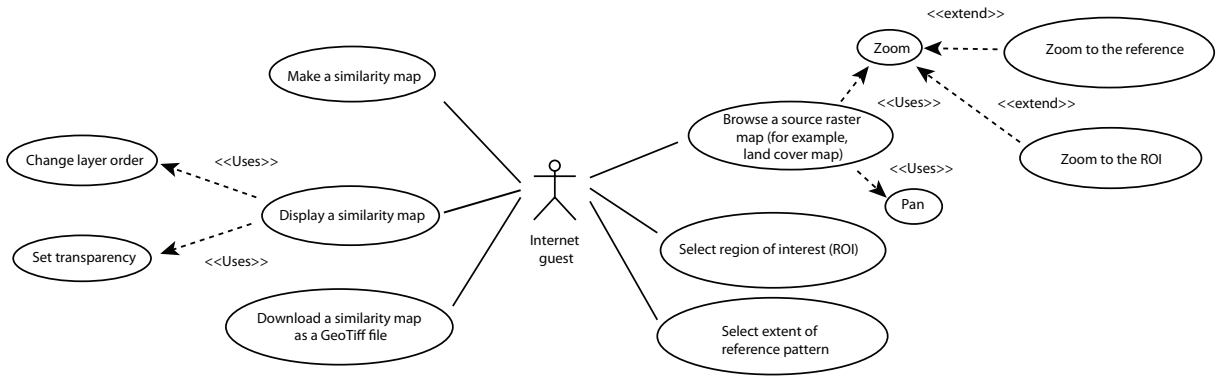

Fig. 4. Use cases diagram for LandEx interface.

similarity map in response to a query. LandEx architecture can be realized on one computer or on several different machines. This makes the system scalable and allows it to fit to load while preserving a fast response to each query.

IV. Examples

As a GeoWeb application, LandEx is a dynamic content intended to be experienced interactively in the web browser. In this section we give four examples of LandEx usage within constraints imposed by a static medium of conventional figures. A reader is encouraged to follow up on these examples by using LandEx. In particular, dependence of similarity map on spatial scale varies with a particular choice of a query but can be readily examined using the tool. Note that these examples are meant to highlight how LandEx can be applied to explore the land cover dataset, but are not intended as actual contributions to featured domains.

A. Cultivated crops with different criss-crossing patterns

One of the 16 land cover classes in NLCD 2006 is the "cultivated crops" class. The spatial extent of this class is shown as a brown color on the NLCD 2006 map (see Fig. 5A). Using existing GIS tools based on SQL-like queries, all cells labeled as "cultivated crops" can be extracted. Extracted region would have a sieve-like topology because areas containing crops are penetrated by roads, building and other features assigned to different land cover classes.

In LandEx, a query identifies regions having similar patterns of classes rather than pixels having the same class. A scene dominated by cultivated crops is characterized not only by that class, but also by a spatial pattern of minor classes reflecting intrusions into cropland. We have selected two scenes completely dominated by the cropland class. The first scene, extracted from the state of North Dakota and shown in Fig. 5C inset, shows cropland criss-crossed by a sparse network of roads. The second scene (of the same size), extracted from the state of Ohio and shown in Fig. 5D inset, shows cropland criss-crossed by a denser network of roads and punctuated by a few very small plots of forest. These queries were processed by LandEx using a coarser spatial scale (see section 2.2) yielding two US-wide similarity maps shown in Fig. 5C and Fig. 5D, respectively.

Examination of the two similarity maps reveals regional differences in the spatial character of cropland across the US. Sparsely criss-crossed croplands (exemplified by the first query) are located mostly in the states of North Dakota, Minnesota, and Iowa, whereas more densely criss-crossed croplands (exemplified by the second query) are mostly restricted to Ohio. Explaining these differences is beyond the scope of this paper. Note that using LandEx it took only a very short time to explore this potential issue.

B. Forest consolidation

Land cover maps have a long history of uses in United States forestry science and management [31]. In particular, analyses of spatial patterns in NLCD were used to report forest fragmentation statistics on a national scale. Forest fragmentation threatens the sustainability of forest interior environments, thereby endangering subordinate ecological attributes and functions. Standard GIS indices, such as Forest Area Density (FAD), have been used to assess spatial distribution of degree of fragmentation [32]. LandEx presents an alternative method of conducting such assessment. An exploration of the spatial distribution of specific forest patterns may be performed by selecting a representative query and analyzing a resultant similarity map. A particularly simple exploration involves submitting a query that represents a region dominated as much as possible by a single class of forest.

Fig. 6 shows the results of two such queries, both using a finer spatial scale (see section 2.2). The first query (shown in Fig. 6A inset) represents a region of consolidated evergreen forest. The resultant similarity map (Fig. 6A) shows the US-wide spatial distribution of similarity to the query, which could be considered as a proxy for the degree of forest fragmentation. High values of similarity (red colors on Fig. 6A) indicate high similarity to the query - regions totally dominated by consolidated evergreen forest. Decreasing values of similarity indicate a lesser similarity to the query (orange and yellow colors on Fig. 6A) - regions only partially occupied by evergreen forest, due most likely to an increased degree of fragmentation. The second query (shown in Fig. 6B inset) represents a region of consolidated deciduous forest. The resultant similarity map (Fig. 6B) could be considered a proxy for US-wide spatial distribution of the degree of deciduous

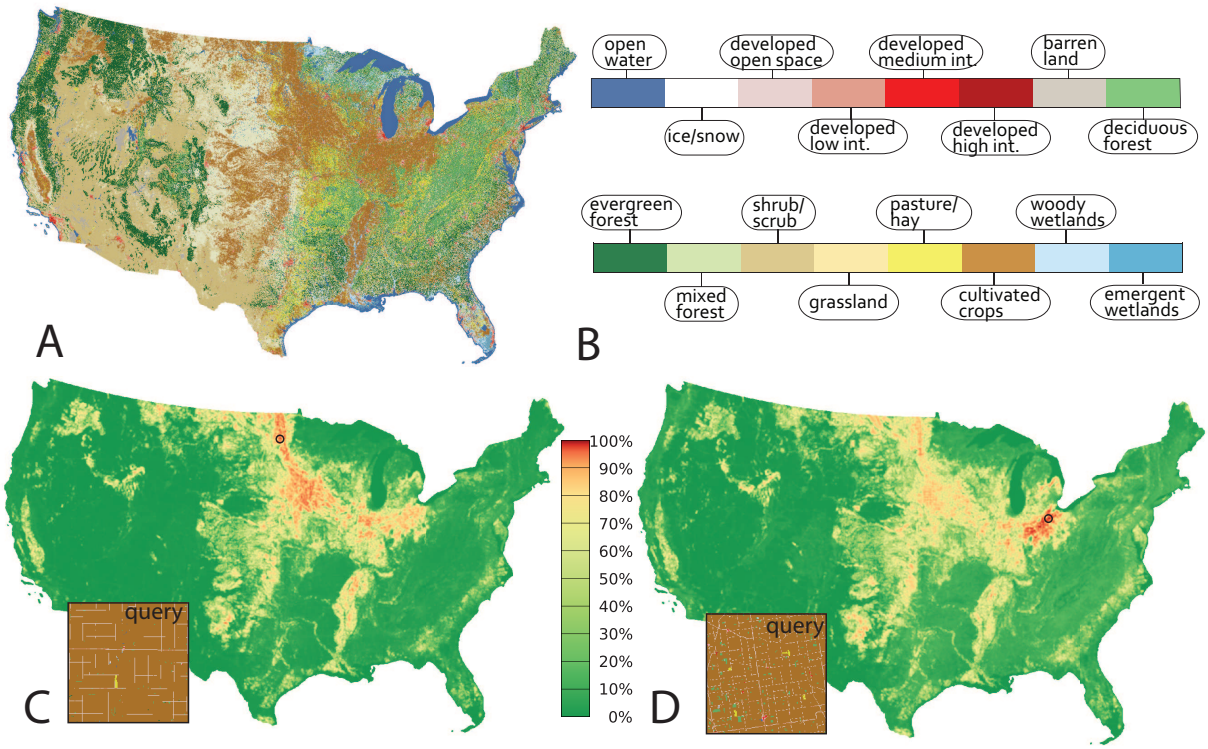

Fig. 5. Similarity maps for croplands with different criss-crossing patterns. (A) The NLCD 2006 map. (B) The legend to the NLCD 2006 map. (C) Similarity map in response to the sparsely criss-crossed cropland query shown in the inset; a circle indicates the location of the query scene. (D) Similarity map in response to the densely criss-crossed cropland query shown in the inset; a circle indicates the location of the query scene. Spatial resolution of similarity maps is 3 km.

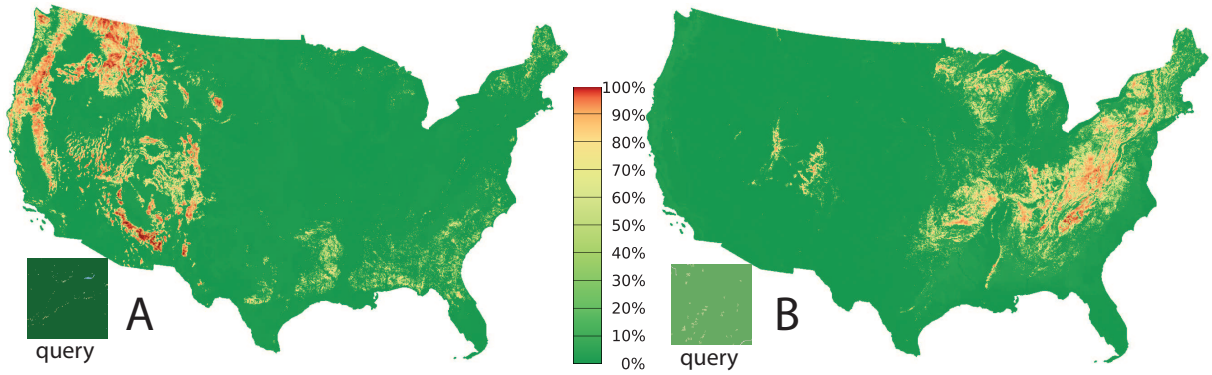

Fig. 6. Similarity maps for queries pertaining to: (A) consolidated evergreen forest, and (B) consolidated deciduous forest. Similarity legend could be used as a proxy for degree of forest segmentation. The Spatial resolution of these similarity maps is 1.5 km.

forest fragmentation. Note that these results were generated by LandEx with minimal effort; the bulk of effort went into finding the representative queries. An extensive exploration of forest patterns can be conducted by running a large number of targeted queries and analyzing their results.

C. Similarity of urban land cover patterns

Urbanization is the process of transforming natural or agricultural landscapes into built-up environments. Thus, in land cover maps, such as the NLCD 2006, urban areas appear as patterns of developed, natural, and cultivated classes. The urban patterns change from one location to another reflecting local differences in the composition of the pre-urbanization landscape and an array of factors related to the actual urbanization process. Quantitative analysis of urban land cover patterns is traditionally performed using landscape indices (for references see [33]). Most studies focus on growing urbanization and thus compare land cover patterns of the same area at different times. Work on comparing patterns of different urban areas at a fixed time step is limited [34] but still based on methodology of landscape indices. LandEx offers a complementary method of studying urbanization patterns. Whereas the landscape indices methodology is most useful for in-depth comparison of two or three a priori selected urban

areas, LandEx enables exploration analysis with the goal of finding all locations featuring urban patterns most similar to a given query.

In order to demonstrate the application of LandEx in such context we have selected a 4.5km $\times$ 4.5km query featuring Clear Lake - a suburb of Houston, Texas and the home of the NASA Johnson Space Center. The search was executed using the finer spatial scale (see section 2.2). Fig. 7 shows the results of our query; although the query was executed for the entire NLCD 2006 dataset, only the Houston metropolitan area is shown. Fig. 7A shows the fragment of the NLCD 2006 restricted to the Houston metropolitan area. Fig. 7B shows the similarity map restricted to the same area. The query and the three other selected regions, labeled C, D, and E are highlighted on the similarity map. The closeups of these regions are also shown in Fig. 7. Regions C and D are examples of locations having a pattern of land cover similar to the query. The region E covers downtown Houston and is clearly not similar to the query. This is because the downtown is dominated by the class labeled as "developed, high intensity" which is rare in the query and the other two regions (all suburbia of Houston). The similarity map gives the geographical distribution of the parts of Houston most similar to Clear Lake from the point of view of the land cover pattern. The areas similar to Clear Lake are all suburbia, but not all of Houston suburbia are similar. The actual analysis of factors responsible for the observed distribution of similarity is beyond the scope of this paper.

D. Rural land cover patterns featuring forest and pasture

For the last example we have selected a 15km $\times$ 15km query located in the Monroe county, Kentucky near the border with Tennessee. This scene (shown in the right inset in Fig. 8A) is dominated by deciduous forest (shown in green) and pasture (shown in yellow) with small addition of cultivated crop class (shown in brown). The two dominant classes in the scene form a pattern of uniform mixture with a characteristic length scale that must reflect natural conditions, land management policies, economic conditions, and societal structures. The search was executed using the coarser spatial scale (see section 2.2). Fig. 8A shows the resultant similarity map restricted to the Kentucky - Tennessee - Missouri area. The entire similarity map is shown in the left inset in Fig. 8A, it shows that most locations with high values of similarity are spatially restricted to the area shown in the main figure. Highly similar patterns occur nearby the location of the query, but also in extended portions of Missouri and Tennessee.

Fig. 8B shows a closeup of the NLCD 2006 map in the immediate vicinity of the query location, whereas Fig. 8C shows the similarity map restricted to the same area. The location of the query is highlighted on both maps. Comparison of the two maps shows in detail how the degree of similarity changes with changing patterns of landscape. Recall that both the query and each local scene have dimensions five times the dimension of a cell in which the value of similarity is stored (the smallest granulation in Fig. 8C).

V. Discussion and future directions

We have developed LandEx - a qualitatively new tool for exploring patterns in large geospatial datasets. Until now geospatial data could only be parsed cell-by-cell using SQL-like queries. Although cell-by-cell comparison of remotely sensed data is useful for local purposes, pattern-by pattern comparison is much preferable for regional-scale or larger-scale assessments because it addresses the question of structural, and thus semantic, similarity. Unlike a cell, pattern has rich enough content to have functional significance for a system, as in the examples considered in section 4. By combining the functionality of computerized maps with the power of pattern recognition algorithms, LandEx makes possible content-based, machine-aided exploration of the entire NLCD. Two elements of LandEx are of particular importance: it's a web-based application available to everyone, and it offers real-time search capabilities. This combination makes it possible to pose questions that are impossible to even formulate without such a tool.

Like all content-based retrieval systems LandEx needs to assign a single number to express objectively but very concisely a degree of similarity between two scene-patterns. Note that this is different from the case when in-depth comparison of two scene-patterns is needed. In such case a lengthened description of similarity is preferable at the price of reduced objectivity. For example, several landscape indices, pertaining to different features of patterns, can be calculated objectively, but the overall degree of similarity needs to be set by a human analyst by weighing the relative importance of these indices. As there is a large number of ways to concisely characterize a pattern, there is also a large number of possible similarity measures [27], but only one (considered to be "well performing" in the context of CBIR) is currently implemented in LandEx. Future work will focus on implementing different measures and assessing their relative performance. Objective evaluation of tools like LandEx is difficult because the notion of "ground truth" cannot be defined. Evaluating performance of LandEx is not unlike evaluating performance of different web browsers - the result depends on the individual preferences of an evaluator. Thus, our future work on the assessment of LandEx performance must be based on the idea of crowdsourcing [35] and aimed at minimizing "a semantic gap" [23] - a difference between similarity as calculated by LandEx and that as perceived by a typical human analyst.

LandEx technology is easily applicable to datasets other than the NLCD 2006. It could be immediately applied to the NLCD 1992 and 2001 given sufficient memory on the server. It can be also applied to the raster version of the CORINE dataset with minimal modification stemming from the larger number of classes and a different map projection. More extensive modification is necessary in order to provide pattern search capabilities to the MODIS 500 m/cell global land cover dataset (MCD12Q10). Global extent of MODIS data requires additional steps to ensure that local tiles have the same size and shape as the query tile. Finally, pattern-based comparison of scenes can be applied to the large-scale assessment of land cover dynamics. Whereas LandEx

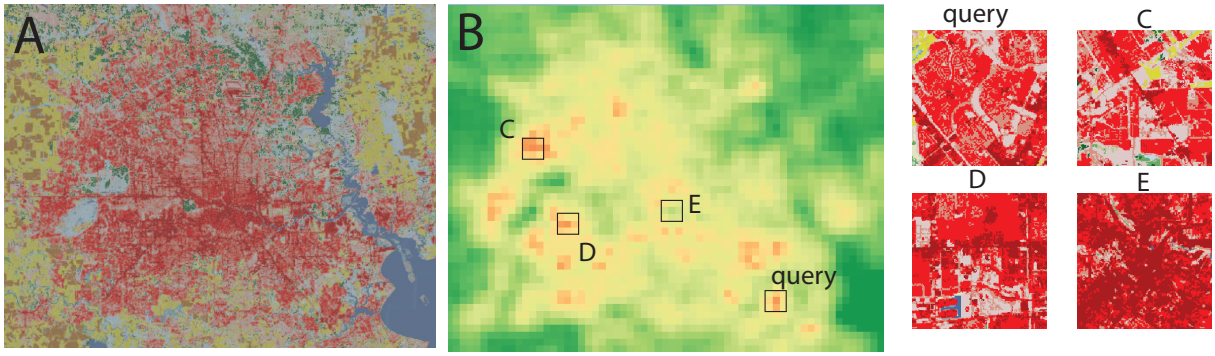

Fig. 7. Similarity map for land cover pattern in Clear Lake, Texas. (A) The NLCD 2006 map of Houston metropolitan area. (B) The similarity map to Clear Lake land cover pattern. Squares indicate locations of query and three other locations described in the main text. Land cover maps of highlighted locations are shown to the right. Spatial resolution of similarity map is 1.5 km. For legends to NLCD 2006 and the similarity values see Fig. 5.

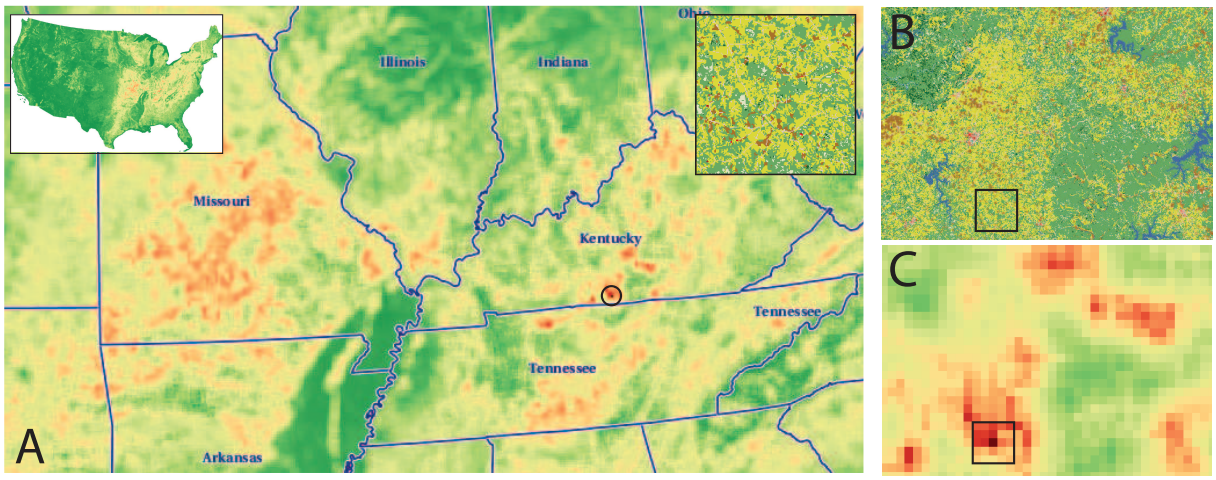

Fig. 8. Similarity map for land cover pattern in the Monroe County, Kentucky. (A) A closeup of similarity map into the Kentucky - Tennessee - Missouri area; circle indicates location of the query. Insets show the entire similarity map (left) and the map of the query (right). (B)/(C) The NLCD 2006 map /similarity map restricted to an immediate vicinity of the query. In maps (B) and (C) a rectangle indicates the location of the query. For legends to NLCD 2006 and the similarity values see Fig. 5.

compares a single query to all other tiles taken from the same dataset, its change-oriented variant would compare all pairs of tiles taken from the same location but at different time steps. Such a tool will find changes in pattern motifs rather than cell-by-cell changes which may be a preferable approach to the global-scale assessment of land cover dynamics. Note that all available datasets have limited classification accuracy. As LandEx uses each dataset "as is" its results (queries or change analysis) can only be as good as the input datasets.

Acknowledgment

This work was supported in part by the National Science Foundation under Grant BCS-1147702 and by the University of Cincinnati Space exploration Institute.

References

[1] M. Datcu, A. Pellizzari, H. Daschiel, M. Quartulli, and K. Seidel. Advanced value adding to metric resolution sar data: Information mining. In *Proc. 4th Eur. Conf. Synthetic Aperature Radar (EUSAR 2002)*, 2002.

[2] T. Gevers and A. W. Smeulders. Content-based image retrieval: An overview. In G. M. S. B. Kang, editor, *Emerging Topics in Computer Vision*, chapter 8, pages 333–384. Upper Saddle River, NJ: Prentice-Hall, 2004.

[3] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys*, 40:1–60, 2008.

[4] M. Lew, N. Sebe, C. Lifi, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput., Commun., Appl.,*, 2(1):1–19, 2006.

[5] M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P. G. Marchetti, and S. D'Elia. Information mining in remote sensing image archives: System concepts. *IEEE Trans. Geosci. Remote Sens.*, 42(12):2923–2936, 2003.

[6] H. Daschiel and M. Datcu. Information mining in remote sensing image archives: System evaluation. *IEEE Trans. Geosci. Remote Sens.*, 43(1):188–199, 2005.

[7] J. Li and R. M. Narayanan. Integrated spectral and spatial information mining in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.*, 42(3):673–685, 2004.

[8] C.-R. Shyu, M. Klaric, G. Scott, A. S. Barb, C. Davis, and K. Palaniappan. GeoIRIS: Geospatial information retrieval and indexing systemContent mining, semantics modeling, and complex queries. *IEEE Trans. Geosci. Remote Sens.*, 45(4):839–852, 2007.

[9] L. Gueguen and M. Datcu. A similarity metric for retrieval of compressed objects: Application to mining sattelite images time series.

IEEE Trans. Knowl. Data Eng., 20(4):562–575, 2008.

[10] J. Fry, G. Xian, S. Jin, J. Dewitz, C. Homer, L. Yang, C. Barnes, N. Herold, and J. Wickham. Completion of the 2006 National Land Cover Database for the Conterminous United States. *PE&RS*, 77(9):858–864, 2011.

[11] J. Jasiewicz and T. F. Stepinski. Example-Based Retrieval of Alike Land-Cover Scenes From NLCD2006 Database. *Geoscience and Remote Sensing Letters*, 10:155–159, 2013.

[12] R. G. Pontius. Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. *Photogrammetric Engineering & Remote Sensing*, 68(10):10411049, 2002.

[13] A. Hagen. Fuzzy set approach to assessing similarity of categorical maps. *Int. J. Geographical Information Science*, 17:235–249, 2003.

[14] T. K. Remmel and F. Csillag. Spectra of coincidences between spatial data sets: Toward hierarchical inferential comparisons of categorical maps. In *GeoComputation International Conference, August 1-3, Ann Arbor, Michigan, USA.*, 2005.

[15] T. K. Remmel and F. Csillag. Mutual information spectra for comparing categorical maps. *International Journal of Remote Sensing*, 27(7):14251452, 2006.

[16] W. W. Hargrove, F. M. Hoffman, and P. F. Hessburg. Mapcurves: a quantitative method for comparing categorical maps. *Journal of Geographical Systems*, 8(2):187–208, 2006.

[17] R.V. O'Neill, J.R. Krummel, R.H. Gardner, G. Sugihara, B. Jackson, D.L. DeAngelis, B.T. Milne, M.G. Turner, B. Zygmunt, S.W. Christensen, V.H. Dale, and R.L. Graham. Indices of landscape pattern. *Landscape Ecology*, 1(3):153–162, 1988.

[18] J. D. Wickham and D. J. Norton. Mapping and analyzing landscape patterns. *Landscape Ecology*, 9(1):7–23, 1994.

[19] T. R. Allen and S. J. Walsh. Spatial and compositional pattern of alpine treeline, Glacier National Park, Montana. *Photogrammetric Engineering & Remote Sensing*, 62(11):1261–1268, 1996.

[20] D. H. Cain, K. Riitters, and K. Orvis. A multi-scale analysis of landscape statistics. *Landscape Ecology*, 12:199212, 1997.

[21] K. H. Riitters, R.V. ONeill, C.T. Hunsaker, J.D. Wickham, D.H. Yankee, S.P. Timmins, K.B. Jones, and B.L. Jackson. A factor analysis of landscape pattern and structure metrics. *Landscape Ecology*, 10:23–39, 1995.

[22] F. Herzog and A. Lausch. Supplementing land-use statistics with landscape metrics: Some methodological considerations. *Environmental Monitoring and Assessment*, 72(1):37–50, 2001.

[23] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.

[24] A. Rosenfeld and J. L. Pfaltz. Sequential operations in digital processing. *J. ACM*, 13:471–494, 1966.

[25] S. Cha. Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307, 2007.

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

[27] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann. Empirical Evaluation of Dissimilarity Measures for Color and Texture. *Computer Vision and Image Understanding*, 84:25–43, 2001.

[28] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379423 & 623656, 1948.

[29] D. R.J. Moore. The Anna Karenina Principle Applied to Ecological Risk Assessments of Multiple Stressors. *Human and Ecological Risk Assessment: An International Journal*, 7(2):231–237, 2001.

[30] M. Neteler and H. Mitasova. *Open source GIS: a GRASS GIS approach*. Springer, New York, third edition edition, 2007.

[31] K. H. Riitters and R. A. Gregory. Applications of national land cover maps in united states forestry. In J. C. Campbell, K. B. Jones, J. H. Smith, and M. T. Koeppe, editors, *North America Land Cover Summit*, pages 97–106. Association of American Geographers, 2008.

[32] K. H. Riitters and J. D. Wickham. Decline of forest interior conditions in the conterminous united states. *Sci. Rep.*, 2:653, 2012.

[33] E. Uuemaa, M. Antrop, J. Roosaare, R. Marja, and U. Mander. Landscape metrics and indices: An overview of their use in landscape research. *Living Rev. Landscape Res.*, 3:1, 2009.

[34] W. C. Hung, Y. C. Chen, and K. S. Cheng. Comparing landcover patterns in Tokyo, Kyoto, and Taipei using ALOS multispectral images. *Landscape and Urban Planning*, 97:132–145, 2010.

[35] D. C. Brabham. Crowdsourcing as a model for problem solving. An introduction and cases. *Convergence: The International Journal of Research into New Media Technologies*, 14:75–90, 2008.

Tomasz Stepinski is the Thomas Jefferson Chair Professor of Space Exploration at the University of Cincinnati. He is a geoscientist with a PhD degree in Applied Mathematics from the University of Arizona. His recent area of research is automated identification, characterization, and classification of landforms and content-based query and retrieval of landscapes. He has established the Space Informatics Lab (http://sil.uc.edu/) at the University of Cincinnati. His E-mail address is stepintz@uc.edu

Pawel Netzel is a mathematician, climatologist, and programmer. He works at the University of Wroclaw, Poland in the Dept. of Climatology and Atmosphere Protection. He promotes and develops Open Source Software for GIS. He is one of the founder-members of OSGeo chapter Poland. His interest includes spatial analyses, especially with applications of AI and adaptative systems, geographical web services, and tele-detection methods of atmospheric boundary layer. His E-mail address is netzelpl@ucmail.uc.edu

Jaroslaw Jasiewicz works at the Adam Mickiewicz University, Faculty of Geography in Poznan, Poland. His research concentrates on the application of computer science methods, such as machine vision, data mining, and search and retrieval, to problems in geoscience and archaeology. He is a contributor of computer codes to the GRASS DEV Team. His E-mail address is jarekj@amu.edu.pl