CLASSIFYING PATTERNS OF LAND COVER USING MUTUAL INFORMATION AND CLUSTERING

Tomasz F. Stepinski

Lunar and Planetary Institute, 3600 Bay Area Blvd., Houston, TX 77058, USA

1. INTRODUCTION

Comparison between two maps of land use/land cover (LULC) is a fundamental task in remote sensing and geospatial data analysis with application to change detection, validation of models, and accuracy assessment [1, 2]. We present a methodology of calculating a similarity measure between every pair of LULC maps in a collection in order to classify them into characteristic LULC patterns. A pattern is a specific composition of LULC categories and their particular spatial arrangement; it represents a higher level abstraction of landscape than a single LULC category. For example, a LULC map of a city provides visual assessment of spatial relations between its constituent categories, but it also defines a pattern - characteristic fingerprint of this particular city in terms of LULC. A collection of different cities can be grouped into classes on the basis of similarities between their patterns. Traditional approaches to quantification of LULC patterns focus on either composition (histogram of LULC categories) or spatial configuration (various pattern indicators). However, for the purpose of classification, it is sufficient to quantify only a *similarity* between two patterns without quantifying patterns themselves. We use a similarity measure [2] that simultaneously takes into account composition and configuration information. First, this similarity measure is applied to every pair of maps in a collection; the output of this step is a similarity matrix that encapsulates pairwise similarities between the maps. Second, we use a combination of clustering and visualization methods to translate the information contained in the similarity matrix into classification of maps. The output of this step is a similarity map that depicts the overall structure of similarities between various maps in the collection, and a classification scheme that groups maps into characteristic LULC-patterns.

2. METHODS

We refer to a LULC map simply as a map and to LULC categories as colors. Distance between the two maps A and B, denoted as $d(A, B)$, is a symmetric function that increases with an increased dissimilarity between A and B. Similarity between maps A and B is given by $\text{sim}(A, B) = 1/(1 + d)$; the range of $\text{sim}(A, B)$ is between $\text{sim} = 0$, where the two maps are as distinct as possible, and $\text{sim} = 1$, where the two maps are identical. Distance between the two maps is calculated using a method [2] which uses normalized mutual information [3] between a stochastic process of selecting a map (or a region in the map) and a stochastic process of selecting a color in the maps to measure $d(A, B)$.

In order to expedite the calculations it is convenient (but not necessary) if both maps have square shapes with a size equal to $s = 2^p$ pixels. This allows to organize the information carried by the pixels into so-called quadtree (Q-tree) data structure. The root of the tree is the entire map. The first four nodes contain data corresponding to the four quadrants (1 or NW, 2 or NE, 3 or SW, 4 or SE) of the map. Subsequent nodes carry data corresponding to regions of the map resulting from recursive subdivision
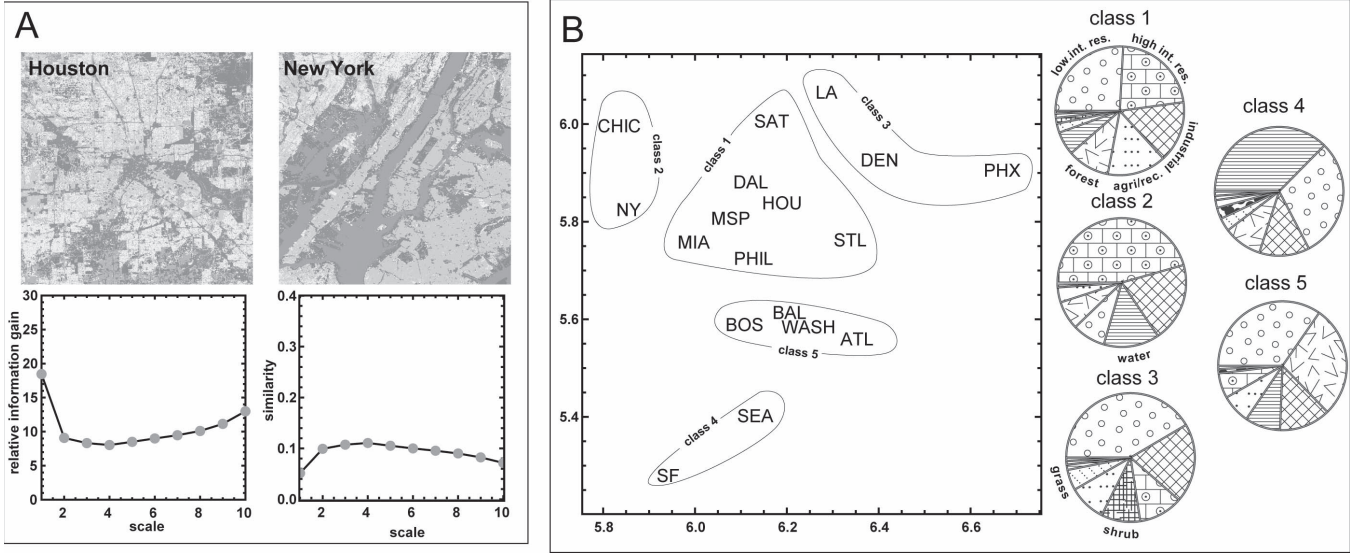
Fig. 1. (A) An example of similarity calculations between two LULC maps labeled as "Houston" and "New York." Dependence on the spatial scale for both, relative information gain (distance) and similarity, are shown. (B) A similarity map of 18 cities; clustering-based classification into 5 classes is indicated by thin encirclements. Pie diagrams show characteristic composition of LULC categories in each class.)

of quadrants into four subquadrants. The terminal nodes (leaves) carry data pertaining to individual pixels. Single pixels are referred to by a list $Y = \{l_1, \ldots, l_p\}$, where $l_i \in \{1, 2, 3, 4\}$, describing a path from the root of the Q-tree to an appropriate leaf. Larger regions (subqadrants) of the map are referred to by shorter paths.

A distribution of colors in two maps is given by a probability function $p(X, Y, Z)$, where X is a map variable ($X = A$ or $X = B$), Y is a spatial location variable (the Q-tree path), and Z is the color variable. Normalized mutual information between $p(X, Y)$ and $p(Z)$ is calculated as a proxy of "distance" between maps A and B. The Q-tree structure facilitates efficient calculations of mutual information between maps at multiple scales through simple adjustment of the length of Q-tree path Y; calculations with length ($Y = 0$) yield distance on the scale of the entire map (scale=1) whereas calculations with length ($Y = p$) yield distance on the pixel scale (scale=$p+1$). The end result of mutual information calculations is a list $D = \{d_1, d_2, \ldots, d_{p+1}\}$ of distances calculated at all the scales corresponding to all levels in the Q-tree structure; those distances may be converted to similarities using the formula given above. Repeating such calculation for all pairs of maps in a collection yields a similarity matrix; a single entry in the matrix is a list of similarities at various scales.

We use two complementary techniques to classify maps on the basis of similarity matrix. The primary technique is the agglomerative clustering. However, clustering does not offer an immediate insight into the overall structure of the map collection; the best number of clusters, their separations, and their within-cluster dispersions are not readily available from the results of a clustering algorithm. In order to address the shortcomings of clustering, we use Sammon's map [4] technique to visualize an overall structure of distances in the entire collection of maps. Clustering and visualization techniques complement each other, using both improves the accuracy of classification.

3. DATA

In order to illustrate our method of classifying landscape patterns we extracted LULC maps for a collection of 18 large metropolitan areas in the United States from the National Land Cover Dataset 1992 (NLCD 1992) [5]. In order to expedite the calculations we degraded the original 30 m/pixel spatial resolution to 60 m/pixel and aggregated the original 21 LULC cat-

egories to only 10 (water, low intensity housing, high intensity housing, commercial/industrial/transportation, rock/sand/clay, forest, shrubland, grassland, agriculture/recreation, and wetland). We refer to the areas in the extracted maps as "cities." Each map has a size 512×512 pixels and is centered at the center of a city. Following cities are included: Houston TX (HOU), Chicago IL, (CHIC), Dallas TX (DAL), Denver CO (DEN), Los Angeles CA (LA), New York NY (NY), Philadelphia PA (PHIL), Seattle WA (SEA), Washington DC (WASH), San Francisco CA (SF), Phoenix AZ (PHX), Atlanta GA (ATL), Boston MA (BOS), Miami FL (MIA), Minneapolis MN (MSP), San Antonio TX (SAN), Saint Louis MO (STL), and Baltimore MD (BAL). We calculated mutual information-based distance and similarity measures for every city-pair in a collection at the scale of the entire map and at $\log_2 512 = 9$ smaller scales.

4. RESULTS

Fig. 1A shows an example of calculating similarity between LULC maps of two cities: HOU and NY. At the coarsest resolution (scale=1) the similarity is relatively small (the information gain is relatively large) due to significant differences in overall composition between the two maps. The similarity increases at spatial scales 2 to 4 indicating that, at these scales, the two cities have patterns that are more alike than the patterns for the entire maps. At finer spatial scales the similarity decreases implying significant differences at small-scale patterns. Similarity map of the entire collection can be calculated at any desired spatial scale by choosing only values of similarities at a specified scale level. The *overall* similarity between cities A and B, that takes into account all scales, can be calculated as, $\mathrm{Sim}(A, B) = \sum_{i=1}^{p+1} V_i \, \mathrm{sim}_i(A, B)$, where $\mathrm{sim}_i(A, B)$ and V_i denote similarities and weights at scale i, respectively. Scale weights are subjective quantities that depends on the specific application.

Fig. 1B shows a similarity map of 18 cities constructed using a portion of similarity matrix corresponding to scale=1; cities located close to each other on the map are similar. The similarity map suggests that 18 cities may be divided into five clusters corresponding to distinct classes of LULC patterns. Agglomerative clustering provides memberships of these classes as indicated by encirclements on the similarity map. Pie diagrams located at the right side of Fig. 1 give characteristic compositions for the classes calculated as average compositions of their members. Class 1 groups cities dominated by residential areas with significant contributions from industrial and agricultural areas. Class 2 is dominated by high intensity residential areas with significant contribution from industrial areas and water. Class 3 groups cities dominated by low intensity residential areas with a significant contribution from industrial areas. Class 4 is dominated by water and low intensity residential areas. Finally, class 5 is dominated by low intensity residential areas and forest. The membership in the classes or the classes themselves may change if the classification is based on similarities corresponding to a different scale level, or a composite similarity that weights contributions from similarities at different spatial scales (see above).

A pair of two cities is characterized by a "relationship" encapsulated by a list of similarities at all scales – a similarity vector. There are 153 pairs of cities in our collection, each pair is characterized by an unique similarity vector. Using the Euclidean distance as a measure of closeness between these vectors we can construct similarity map of relationships between city-pairs and cluster these relationships into groups congregating similar associations. Fig. 2A shows the similarity map of 153 city-pairs. We use agglomerative clustering to cluster all pairs into 5 groups corresponding to 5 different types of relationships. These groups are indicated by encirclements on the similarity map. The membership of city-pairs in the groups is listed. Fig. 2B shows characteristic relationships (similarity vectors shown as curves) for all group calculated as average vectors from their members.

Group 1 corresponds to a "similarity" relationship. Group 5 corresponds to a "dissimilarity" relationship. Group 4 represents relationship similar to that given by group 5, but at somewhat larger levels of similarity; groups 4 and 5 could be combined into a single very large "dissimilarity" group. Group 3 corresponds to relationship that is defined by significantly different overall compositions but also by existence of matching patterns on smaller scales. Finally, group 2 corresponds to relationship defined by somewhat similar compositions at scales 1 and 2 and a gradual decrease of similarities at finer scales.
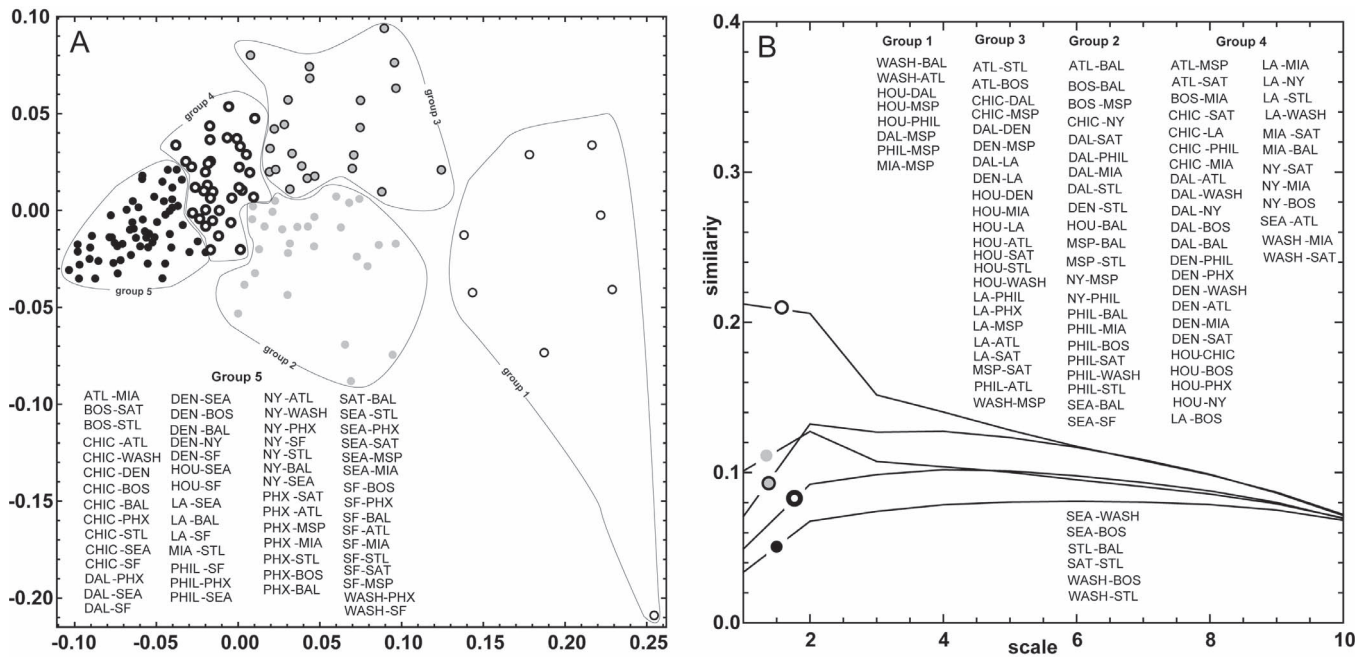
Fig. 2. (A) A similarity map of 153 pairs of cities; clustering-based classification into 5 groups is indicated and the membership in each group is listed. (B) Similarity curves characteristic for each group.)

5. DISCUSSION

Presented method can find a broad application in all comparative studies of landscapes. One such application is to map landscape units larger than individual LULC categories – patterns of specific categories. For example, one of the cities in our collection may be divided into 64 tiles or sectors each having size of 64 × 64 pixels. The collection of 64 sectors can be classified into groups of similar patterns interpreted as downtown, suburbia, industry, etc. Assigning unique labels/colors to these groups results in a meta-map of a city bringing out urban units at higher level of generalization than the LULC categories. The method can also be applied to a collection of LULC maps derived at different times. This will help to catalog various trajectories of landscape change and to identify driving factors affecting landscape dynamics.

6. REFERENCES

[1] Jr. R. G. Pontius, "Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolution," *Photogram. Eng. Remote Sens.*, vol. 68(10), pp. 1041–1049, 2002.

[2] T. K. Remmel and F. Csillag, "Mutual information spectra for comparing categorical maps," *Inter. Journal of Remote Sensing*, vol. 27(7), pp. 1425–1452, 2006.

[3] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 623–656, 1948.

[4] Jr. J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on computers*, vol. C-18(5), pp. 401409, 1968.

[5] J. B. Vogelmann, S.M. Howard, L. Yang, C.R. Larson, B.K. Wylie, and N. Van-Driel, "Completion of the 1990s national land cover data set for the conterminous united states from landsat thematic mapper data and ancillary data sources," *Photogram. Eng. Remote Sens.*, vol. 67, pp. 650–652, 2001.