Ecological Indicators 64 (2016) 20–30

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind

On using landscape metrics for landscape similarity search

Jacek Niesterowicz, Tomasz F. Stepinski *

Space Informatics Lab, Department of Geography, University of Cincinnati, Cincinnati, OH 45221-0131, USA

ARTICLE INFO

Article history:
Received 2 March 2015
Received in revised form 27 October 2015
Accepted 15 December 2015
Available online 7 January 2016

Keywords:
Landscape pattern
Landscape metrics
Similarity search
Regionalization
Similarity measure

ABSTRACT

Landscape similarity search involves finding landscapes from among a large collection that are similar to a query landscape. An example of such collection is a large land cover map subdivided into a grid of smaller local landscapes, a query is a local landscape of interest, and the task is to find other local landscapes within a map which are perceptually similar to the query. Landscape search and the related task of pattern-based regionalization, requires a measure of similarity – a function which quantifies the level of likeness between two landscapes. The standard approach is to use the Euclidean distance between vectors of landscape metrics derived from the two landscapes, but no in-depth analysis of this approach has been conducted. In this paper we investigate the performance of different implementations of the standard similarity measure. Five different implementations are tested against each other and against a control similarity measure based on histograms of class co-occurrence features and the Jensen–Shannon divergence. Testing consists of a series of numerical experiments combined with visual assessments on a set of 400 3 km-scale landscapes. Based on the cases where visual assessment provides definitive answer, we have determined that the standard similarity measure is sensitive to the way landscape metrics are normalized and, additionally, to whether weights aimed at controlling the relative contribution of landscape composition vs. configuration are used. The standard measure achieves the best performance when metrics are normalized using their extreme values extracted from all possible landscapes, not just the landscapes in the given collection, and when weights are assigned so the combined influence of composition metrics on the similarity value equals the combined influence of configuration metrics. We have also determined that the control similarity measure outperforms all implementations of the standard measure.

1. Introduction

Similarity search is a key technology for data scientists with applications in information retrieval, data mining, decision making, event detection, etc. (Zezula, 2012). It can be defined as the retrieval of the "closest" objects to a query in a database. Recently, as the notion of landscape as a mosaic of land cover patches spreads from its original application in ecology to general analysis of land use/land cover (LULC) patterns (Uuemaa et al., 2013), the similarity search is starting to be applied to landscapes. Two types of analyses utilize similarity search: (a) regionalization of LULC maps into landscape pattern types (Wickham and Norton, 1994), and (b) landscape search, where the goal is to identify from among a large set of different landscapes those showing perceptual similarity to a template landscape. It is assumed that two perceptually similar landscapes express closely related meanings, and fulfill similar functions.

Pattern-based regionalization aims at delineation of a pixel-based LULC map into sub-regions characterized by unique, stationary LULC patterns. This is achieved by dividing the entire map into a regular grid of local landscapes (rectangular blocks of pixels interpreted as mosaics of LULC class patches restricted to the interior of a block) which are clustered in data space (using a similarity measure) to discover and delineate landscape pattern types present in the LULC map. From the computational perspective regionalization (Werlen, 2009) is an unsupervised classification resulting in the generalization of the original LULC map into a useful overview across a large area. Cardille and Lambois (2009) applied such a method to the 1992 National Land Cover Dataset (Vogelmann et al., 2001) and identified 17 landscape pattern types summarizing typical patterns (on the spatial scale of 6.5 km) of LULC across the conterminous U.S. Using the Earth Observation for Sustainable Development of Forests land-cover (EOSD) dataset (Wulder et al., 2008). Partington and Cardille (2013) regionalized the forested part of Quebec, Canada into several landscape pattern

* Corresponding author. Tel.: +1 5135563583.
E-mail addresses: niestejk@mail.uc.edu (J. Niesterowicz), stepintz@uc.edu (T.F. Stepinski).

http://dx.doi.org/10.1016/j.ecolind.2015.12.027

types (at a spatial scale of 30 km) and concluded that the method *fulfilled the basic needs of land management across large areas: produced an overview of a large area, while highlighting useful subsets for closer inspection.* Long et al. (2010) regionalized 5.5 million ha of primarily forested land situated within the interior plateau of British Columbia, Canada into six landscape pattern types (at a spatial scale of 1 km) using also the EOSD data.

Unlike a regionalization task, a landscape search task relies on the principles of supervised classification. Given a local landscape (or a set of landscapes) of interest it discovers other local landscapes having similar structures. Andrew et al. (2012) utilized landscape search to identify *de facto* protected areas in boreal Canada. They used structures (patterns of land cover classes) of local landscapes in formally protected areas as a training set for finding local landscapes with similar structures in areas which were not formally protected but presently inaccessible. The identified areas are valuable candidates for protected area expansion. In a similar study Cardille et al. (2012) identified representative local landscapes located in Ontario's parks and protected areas and searched for similar landscapes in parts of Ontario which are currently unprotected. Their goal was to select some of the identified areas for long-term monitoring to establish benefits of protection. In ecology, Dilts et al. (2010) recognized the need for a landscape search in order to identify control sites for sites that will experience treatment (for example, a road construction or forest harvesting). In epidemiology, Roux et al. (2011) used landscape search to infer the occurrence of Chagas disease. They identified landscapes structurally similar to those where the disease was known to be present to enable closer inspection for the possible presence of the disease.

The denominator of all these investigations is that they all incorporate knowledge discovery and the concept of holistic perceptual measure of similarity between landscapes. A holistic perceptual similarity between two landscapes (hereafter referred to as a similarity) is a function that gives a non-negative number to each pair of landscapes to define a notion of an overall likeness or sameness between them without regard to minute differences and spatial orientation. Thus, it assesses, in a single number, a degree to which two landscapes have similar utility or function. A large number of different similarity measures have been proposed in data science literature (Cha, 2007). This is because no single similarity measure is appropriate and effective to all problems, instead the most appropriate similarity measure needs to be matched to the data and the analytic task at hand. However, all the aforementioned studies on landscape regionalization and search used the same, standard approach to calculating a similarity measure. This standard approach consists of representing landscapes as feature vectors of landscape metrics (LMs) (Haines-Young and Chopping, 1996; Herzog and Lausch, 2001) and calculating the Euclidean distance between feature vectors to quantify similarity between two landscapes. To the best of our knowledge no study has been conducted which investigates the appropriateness and effectiveness of this approach. As similarity search-based analyses of landscapes are becoming more frequent, it is important to use an optimal measure of similarity. The aim of this paper is to investigate how different implementations of the standard landscape similarity measure affect the results, and to recommend its best implementation. An additional aim is to compare a standard similarity measure with a measure based on histograms of class co-occurrence features and the Jensen-Shannon divergence.

2. Landscape similarity

Following McGarigal et al. (2002) we use the term "landscape" as ... *an interacting mosaic of patches relevant to the phenomenon under consideration*. In this paper LULC patterns are referred to as landscapes and the sought after LULC pattern types are referred to as landscape pattern types. Thus, the landscape we consider is the LULC raster which has K nominal labels $c_1, \ldots, c_K$ describing K land cover classes. The entire spatial extent of the LULC raster map is referred to as a region. The region is subdivided (without overlap) into a lattice of local landscapes (LL). Thus, from the data science perspective, a region is a spatial database of LLs. Landscape search can be thought of as a special case of similarity search – the retrieval of "closest" objects (LLs) to a query (a selected LL) in the database. Regionalization can be thought of as a clustering of all objects (LLs) in database. A specific LL (denoted by $\mathcal{A}$) is a $n \times n$ block of pixels. The size n sets the spatial scale over which the pattern of LULC classes is defined as a local landscape. A landscape similarity measure requires two elements: (a) a mathematical description of the landscape (called a signature) and (b) a similarity function which takes two signatures as arguments and returns the value of similarity between the landscapes.

2.1. Calculating landscape dissimilarity using landscape metrics

In working with landscapes the standard practice is to use a vector of LMs as a landscape signature and the Euclidean distance as a similarity function. Note that *distance*, which assesses the degree of "unlikeness" between two patterns, is the opposite of similarity and is better referred to as dissimilarity in the present context. We use the notions of similarity and dissimilarity interchangeably as they are easily convertible.

LMs are algorithms that quantify the specific spatial characteristics of a landscape pattern. A large number of different metrics characterizing individual patches, classes of patches, and entire landscape mosaic have been developed and collected in a single computer program FRAGSTATS (McGarigal et al., 2002). A patch is a contiguous group of same-class pixels. In a standard approach a signature of landscape $\mathcal{A}$ is a vector $(a_1, \ldots, a_N)$ consisting of the values of N different LMs calculated for $\mathcal{A}$ and the dissimilarity function is the Euclidean distance between two LLs $\mathcal{A}$ and $\mathcal{B}$

$$d_E(\mathcal{A}, \mathcal{B}) = \sqrt{(\mu_1(a_1 - b_1)^2 + \cdots + \mu_N(a_N - b_N)^2)} \qquad (1)$$

where $(a_1, \ldots, a_N)$ and $(b_1, \ldots, b_N)$ are vectors of LMs calculated from $\mathcal{A}$ and $\mathcal{B}$, respectively, and $(\mu_1, \ldots, \mu_N)$ are the weights to reflect the relative importance of a given LM to the overall value of dissimilarity. There are a number of issues that arise with the standard approach: (a) Which LMs should be selected for a landscape signature? (b) How should LMs in the signature vector be normalized or standardized if at all? (c) What values of weights should be assigned, if any?

2.1.1. Selection of landscape metrics

Compositional metrics (Gustafson, 1998) need to be included in the signature as the composition of a landscape is its primary characteristic. Only landscape-level configuration LMs should also be included in a landscape signature because only landscape-level LMs can be calculated for all LLs regardless of their individual composition. Ideally, a landscape signature should consist of the values of LMs that are independent and together describe adequately the character of the pattern for all possible patterns in a region. Cushman et al. (2008) identified seven landscape structure components – linear combinations of LMs obtained using the principal components analysis (PCA) – that were independent and universal, at least on a set of 531 landscapes (at a ~7 km length scale) from across three different regions in the U.S. Conceivably, those components could be good candidates for landscape signatures but because they have been established on the basis of a single scale landscape and over a limited number of all possible U.S. landscapes they may not apply to all landscapes across the entire United States.