Applied Geography 75 (2016) 60–70

Contents lists available at ScienceDirect

Applied Geography

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

Characterising landscape variation through spatial folksonomies

Curdin Derungs [a, b, *], Ross S. Purves [a, **]

[a] Department of Geography, University of Zurich, Switzerland
[b] URPP Language and Space, University of Zurich, Switzerland

ARTICLE INFO

Article history:
Received 11 July 2015
Received in revised form
3 August 2016
Accepted 8 August 2016
Available online 20 August 2016

Keywords:
Land cover classification
User generated content
Text
Geographic information retrieval
Ecosystem services
CORINE
Spatial folksonomy

ABSTRACT

Describing current, past and future landscapes for inventory and policy making purposes requires classifications capturing variation in, for example, land use and land cover. Typical land cover classifications for such purposes result from a top-down process and rely on expert conceptualisations, and thus provide limited space for incorporating more widely held views of key landscape elements. In this paper we introduce the notion of spatial folksonomies, which we define as a tuple linking a vocabulary of landscape terms through authors and resources to locations. We demonstrate how spatial folksonomies can automatically be created for Switzerland using two text corpora: the Swiss Alpine Club's yearbook for the past 150 years and user generated content from a website describing a wide range of outdoor activities. The spatial folksonomies capture variation in space of the use of nouns describing 96 natural landscape terms (e.g. ridge, forest, mountain, etc.) and allow us to characterise regions and compute similarities. We compare our spatial folksonomies to two traditional land cover/land use classifications (CORINE and Arealstatistik) and demonstrate that despite their very different sources, the approaches capture landscape variation in broadly similar ways. However, our spatial folksonomies provide new insights into how landscapes are described, through for example variation in space, time and through the prism of different activities. We argue that our spatial folksonomies are a novel way of capturing variation closer to the bottom-up understandings of landscape for instance required to describe cultural ecosystem services.

1. Introduction and background

Spatially explicit geographic information describing land use, land cover and landscapes[1] is today indispensable at research and policy levels, not only for inventory purposes, but also in the quantification and modelling of past (e.g. Feranec, Jaffrain, Soukup, & Hazeu, 2010; Gibbs & Salmon, 2015) and projected future changes (e.g. Feddema et al., 2005; Price et al., 2015). Thus, the European Environment Agency maintains CORINE land cover data arguing:

> "If our environment and natural heritage are to be properly managed, decision-makers need to be provided with both an overview of existing knowledge, and information which is as complete and up-to-date as possible on changes in certain features of the biosphere."(EEA-ETC, 1994, p. 3)

CORINE, is based on the interpretation of imagery, and compiled using an expert classification schema for allocating areas to land cover classes. This allows comparison between regions using a single, shared, vocabulary of terms. However, the resulting approach can only be performed by experts and the vocabulary thus produced can be seen as a top-down process. Despite the complex process of negotiating an agreed classification, inventories are challenged by issues relating to not only technologies (e.g. differences between sensors), but also ontologies (what exactly is a mountain or a forest?) and their embedding in societies within which particular landscapes, land uses and land covers are differently valued (Comber, Fisher, & Wadsworth, 2005). Furthermore, typical land cover and landscape classifications have limited meaning for average citizens (e.g. *transitional woodland shrub* is a typical CORINE class) despite the recognised need to involve citizens in policy:

* Corresponding author. Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland.
** Corresponding author.
E-mail addresses: curdin.derungs@geo.uzh.ch (C. Derungs), ross.purves@geo.uzh.ch (R.S. Purves).

[1] We deliberately include all three concepts here.

http://dx.doi.org/10.1016/j.apgeog.2016.08.005

> "A landscape policy which involved only experts and administrators, who themselves are often specialists, would result in landscapes that were imposed on the public, just as in the days when landscape was produced by and for an elite." (Prieur et al., 2006, p. 28, p. 28)

If landscape inventories are to be meaningful and useful as tools in exploring policy from the perspective of citizens, classifications need to be linked to ways in which individuals and cultural groups share conceptualisations of landscapes (Prieur et al., 2006). One current set of approaches to incorporating such non-expert conceptualisations of land use and land cover involves the use of crowdsourcing methods, where individuals can, classify particular locations (e.g. Perger et al., 2012) or participate in evaluating data (e.g. Fritz et al., 2009). Research in other geographic fields has indicated the potential of user generated content (UGC), such as images and other sources labelled by individuals (e.g. in Flickr), in deriving information about how individuals name locations (e.g. Hollenstein & Purves, 2010) or documenting forest fires (Spinsanti & Ostermann, 2013). In parallel, information scientists have used UGC to develop folksonomies, defined by Hotho, Jäschke, Schmitz, and Stumme (2006) as follows:

> "… 'folksonomy' is a blend of the words 'taxonomy' and 'folk', and stands for conceptual structures created by people …" (p.411).

Many of those producing folksonomies argue that the folk-centred nature of the information contrasts with expert knowledge often used in more formal data structures, such as ontologies, and provides access to more bottom-up conceptualisations (e.g. Gruber, 2007). Typically, folksonomies are considered to be formed from a triple of *users* annotating *resources* with *tags* (Winget, 2006). The bottom-up nature of folksonomies is argued to result from their emergent nature, whereby tags used frequently by many users suggest shared conceptualisations (Hollenstein & Purves, 2010; Winget, 2006). Since, for example, individual resources or users can be associated with weighted vectors of tags, it is also possible to calculate similarity between resources or users using a range of similarity measures (Cantador, Bellogín, & Vallet, 2010).

In parallel to developments focussing on UGC, the availability of digitized texts in general has significantly increased in recent years. Thus, Google claims to have digitized and made available some 6% of books ever published, resulting in an n-gram corpus of nearly half a trillion words in English (Michel et al., 2011). Clearly, if such texts can be explicitly linked to space, then it is not only possible to explore how a particular theme is discussed over time, but also where. This linking process forms the core of methods in Geographic Information Retrieval (Purves & Jones, 2006) focussing on firstly, identifying references to named places in a text; secondly, disambiguating such references to a single geographic location and, thirdly, associating these locations with the text passages for which they are relevant. In previous work we developed a set of methods specifically designed to perform this task for a mountaineering corpus (Derungs & Purves, 2014).

Together, these developments motivate our work in this paper. From the above it is clear that there is a need for landscape, land use and land cover classifications which are closer to everyday conceptualisations, and thus better reflect bottom-up conceptualisations. Equally, if such classifications are to be produced, it seems reasonable to expect that they will vary in space, and thus we introduce the notion of a *spatial folksonomy* which we define as a tuple linking a *vocabulary of terms* through *authors* and *resources* to *locations*.

In the following, we argue that a spatial folksonomy can be created by analysing not simply individual atoms of UGC (such as tags describing images associated with locations), but rather through processing and analysing rich natural language descriptions and associating information contained in such descriptions with locations. We do so using methods introduced in previous work (Derungs & Purves, 2014), and generate two spatial folksonomies for Switzerland, focussing on mountainous regions. Furthermore, we seek to demonstrate that these spatial folksonomies provide complementary, but not discordant, perspectives with respect to traditional data sources such as CORINE. Specifically, we set out to investigate the following three research questions:

> RQ1: How can we automatically and reproducibly produce a spatial folksonomy of landscape terms?
> RQ2: How does such a spatial folksonomy compare to more traditional landscape characterisations, such as CORINE?
> RQ3: How can a spatial folksonomy be used to enable discussions on landscape, land use and land cover?

2. Data

To build our spatial folksonomy we used two contrasting digitized corpora, Text + Berg (Volk, Bubenhofer, Althaus, & Bangerter, 2010) and HIKR (www.hikr.org) both of which contain reports pertaining to mountaineering activities in Switzerland. These corpora were selected for three reasons:

1. They cover the same region (Switzerland) and broadly similar activities (mountaineering), and their contents were authored by large numbers of contributors.
2. They have very different historical backgrounds: Text + Berg is the digitized yearbook of the Swiss Alpine club, dating back to 1864, while HIKR is a typical Web2.0 resource containing reports on mountaineering trips dating back to 2003.
3. Finally, and most importantly, we believe that these two corpora are sufficiently rich and varied such that they can provide us with an emergent view of natural feature terms used to describe land cover, land use and landscapes – an essential property for our spatial folksonomy.

For comparative purposes we used two contrasting datasets, CORINE, a European land cover dataset and *Arealstatistik* a Swiss land use dataset. Key features of each of the four data sets relevant to our study are now described in turn.

The Text + Berg corpus contains 150 Alpine yearbooks dating from 1864 to the present, and consists of articles relating to mountaineering, climbing or hiking and other material of interest to members of the Swiss Alpine Club. The yearbooks are edited volumes, published initially in a mixture of German, French and occasionally Italian and laterally in parallel in all three languages. Earlier versions of the yearbook underwent a rigorous editorial process and were contributed to by a relatively select number of authors with a small, specialised, readership. Newer volumes have a broader authorship and are published for a very wide audience with approximately 130,000 readers. We have available pre-processed texts on which part-of-speech tagging in German has been performed (Sennrich, Schneider, Volk, & Warin, 2009). In total, more than 10,000 individual articles were processed, with an average length of 1500 words.

HIKR is a UGC corpus where users describe outdoor activities, such as mountaineering, climbing or hiking trips. The descriptions have average lengths of around 500 words, with only 3% having