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
0143-6228/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

> “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 preprocessed 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 more than 1000. We identified and selected some 25,000 descriptions written in German for our analysis, which is about half of the corpus. As is typical in UGC, descriptions are not equally contributed by authors, and some 90% of all descriptions are written by only 1% of the users (n = 10,000, of whom many are *passive* users). The earliest articles date back to 2003. The number of articles added per year steadily increased until 2011 since when it has stagnated.

We used the most recent available version of CORINE dating from 2006. Data compilation is the responsibility of individual member states, but follows a standardized process and classification schema, consisting of three hierarchical levels and 44 individual classes at the most detailed scale of which 29 occur in Switzerland (Steinmeier, 2013). CORINE contains land cover data for units captured at a minimum mapping size of 25ha.

Arealstatistik was first introduced in the 1980s as a Swiss federal product. Data compilation covers all of Switzerland every 12 years. The last complete version of Arealstatistik, which we use, is from 2004. Land cover is captured on 100 m resolution lattice, using a classification schema with 72 classes (Hotz & Weibel, 2005).

3. Creating and comparing spatial folksonomies

3.1. Populating spatial folksonomies

In our data as we retrieve it from the written corpora, we identify the following elements and dependencies: authors, who write contributions, use landscape terms to describe natural landscapes at particular locations. To formally capture these dependencies we broadly follow Cantador et al., (2010)'s definition of a folksonomy, adding the notion of location such that a spatial folksonomy F is defined as a list of elements (i.e. a tuple) F = {T,U,R,L,A}, where T = {t_1, …,t_L} is a weighted set of terms capturing the spatial folksonomy's vocabulary, U = {u_1, …,u_M}, R = {r_1, …,r_N} and L = {l_1, …,l_P} are the set of authors (users) U who use terms to describe resources R at locations L and A = {(t_l,u_m,r_n,l_p)}∈T × U × R × L is a set of assignments linking terms to resources and locations through authors. In this paper we simplify our analysis by neglecting individual users, and thus reduce our spatial folksonomy to the following tuple F = {T,R,L,A}.

To populate our spatial folksonomy it is necessary to find assignments of terms from our vocabulary in individual resources and link these to locations. Our approach to this process was described in detail in Derungs and Purves (2014), and here we briefly summarise the key steps of the process:

1. Four individuals annotated natural landscape feature terms in a list containing the 1500 most frequently used nouns in Text + Berg. As a result, 96 terms are judged to represent natural features by at least three out of the four annotators. The list of terms can be found in the appendix of Derungs and Purves (2014).
2. Natural landscape features from the agreed vocabulary of 96 terms are counted for each resource to produce a term frequency vector.
3. Toponyms are identified and disambiguated in individual resources to produce a set of locations for each resource.
4. For every 10 km grid cell covering Switzerland, we check whether more than 20% of the locations associated with a single resource are contained by the cell. If this is the case, the term frequency vector associated with that resource is added to the term frequencies associated with the cell.
5. For every 10 km grid cell we calculate tf-idf weights, where tf is the term frequency for an individual natural landscape feature within a given grid cell and idf is the inverse of the log of the frequency of a term over the whole corpus.

The final product, a spatial folksonomy is thus a tf-idf weighted term vector for 96 natural features, associated with individual grid cells.

3.2. Comparing the content of spatial folksonomies

The content of the two spatial folksonomies were compared against each other and, in a second step, with two land cover classifications, namely Arealstatistik and CORINE. In both cases, we compared how the same regions, of regular 10 × 10 km size (i.e. grid cells), are described by weighted term vectors (spatial folksonomies) and land cover classes (land cover classifications).

3.2.1. Comparing spatial folksonomies

The two spatial folksonomies have identical data structures. Similarity between these two term vectors can thus be calculated as a correlation value for individual cells. This procedure allows us to make pairwise comparisons between 10 × 10 km cells in Switzerland and derive a quantitative result that can be cartographically represented in two ways. Firstly, for a given location (e.g. the cell containing the Matterhorn) we can calculate similarity to every other grid cell in the corpus, and thus explore whether seemingly similar landscapes are described using similar term vectors. Secondly, we can compare corpora by calculating the correlations between cells at identical locations, thus exploring where more (or less) similar terms are used to describe locations.

3.2.2. Comparing spatial folksonomies with land cover classifications

Comparing land cover and land use classifications with our spatial folksonomies is more complex since both spatial resolutions and classifications differ. We first aggregated the content of both land cover classifications to the 10 × 10 km resolution grid of our spatial folksonomy generating a frequency distribution of the respective Arealstatistik and CORINE classes for each grid cell. After this step a grid cell was either represented by the tf-idf weighted term vectors for 96 natural features from the two spatial folksonomies or analogous frequency distributions from the land cover/land use data from CORINE and Arealstatistik respectively.

A first comparison between each of the two spatial folksonomies and the two land cover/land use data sets was conducted by focussing on the diversity or coverage of each grid cell. Coverage describes the relative number of natural feature terms or classes retrieved for each grid cell. If for instance in the Arealstatistik four land use classes are used to classify the content of a grid cell, its coverage is 6% (4 of 72 classes are represented).

Quantitative similarity comparisons between the heterogeneous cell contents could be computed by treating each of these representations as a simple term vector and calculating an appropriate similarity measure. However, such abstract measures (though useful for given applications) lack interpretability, and here we compare the 15 most prominent/frequent terms for a set of four predefined and heterogeneous regions and discuss the major differences between the contents of spatial folksonomies, as compared to land cover/land use classifications.

4. Results and interpretation

We first discuss the properties of two spatial folksonomies populated by information from the Text + Berg and the HIKR corpus, focussing on how the contents of the two folksonomies, qualitatively and quantitatively, related to each other. In a second step, we compare the spatial folksonomies to administrative land cover classifications in terms of coverage and content. Finally, we stratify the two spatial folksonomies topically (HIKR) and temporally (Text + Berg), to discuss potential application areas of spatial folksonomies.

Fig. 1 contains labels for approximate regions in Switzerland that are used in the following interpretation of the results.

4.1. Properties of the spatial folksonomies

Table 1 summarises key properties of the two spatial folksonomies. Text + Berg contains twice as many tokens as HIKR and three times as many toponyms per article. The size of the corpus is also reflected in the amount of available information at the level of individual grid cells, with Text + Berg having both more natural features per grid cell and more unique natural features per grid cell.

However, although Text + Berg captures more natural features per grid cell, the pattern of coverage is comparable across the two spatial folksonomies (Fig. 2). Both spatial folksonomies are dominated by the Swiss Alps, as one would expect. In HIKR a slight shift to the north indicates the relative importance of the Pre-Alps as a hiking destination.

Fig. 3 compares the most prominent natural feature per cell (according to tf-idf values), occurring in more than 10 cells overall, for the two spatial folksonomies. Five natural features (*gletscher* (glacier), *grat* (ridge), *pass* (pass), *see* (lake) and *wald* (forest)) are shared between the two folksonomies, while the three terms more prominent in HIKR (*gipfel* (summit), *pfad* (path) and *schlucht* (gorge)) appear to be more relevant to descriptions of hikes (e.g. *summits* are usually reached, *paths* are taken and *gorges* are a particularly important feature affording access in the Jura mountains in north-western Switzerland). Finally, forest (*wald*) is clearly an important feature for HIKR users in the populated Mittelland. Text + Berg is somewhat more diverse, but it is also clear that, even when only exploring the most prominent natural feature per cell, similarities are present in the descriptions such as for glacier landscapes (*gletscher*) in the Bernese Oberland and the Valais. It is also important to note that the features mapped here are simply the highest tf-idf values from a term vector containing up to 96 values per cell.

An important strength of our spatial folksonomy approach over existing classifications is the richness of the complete term vectors, and in Fig. 4 we visualize the 15 most prominent natural features as spatial word clouds (Ahern, Naaman, Nair, & Yang, 2007), for eight selected grid cells representing diverse landscape types.

Mountain landscapes, such as the Matterhorn, Salbit, Finsteraarhorn or the Bernina regions are often associated with glaciers (*gletscher*) (c.f. Fig. 3). However, additional features, allow us to identify regional particularities, such as the salience of Finsteraarhorn's peak (*spitze, horn*, *grat*) or the ridges providing access to the summit of the Salbit (*süd-, ost- and westgrat*). The regions of Thun and Uetliberg are both characterised by gentle landscapes with forests (*wald*), trees (*baum*), lakes (*see*) and rivers (*fluss*). Waterfall (*wasserfall*), however, which is only listed for Thun, correlating with the physical presence of waterfalls in this regions, is not contained in the descriptions of Uetliberg, correctly reflecting the lack of waterfalls. Importantly, our methods do not discriminate between descriptions of *being in* or *seeing* a location. Thus, for example, the prominence of glacier (*gletscher*) for the HIKR data from Bernina may well indicate the visual salience of this mountain, and its glaciers, from long distances.

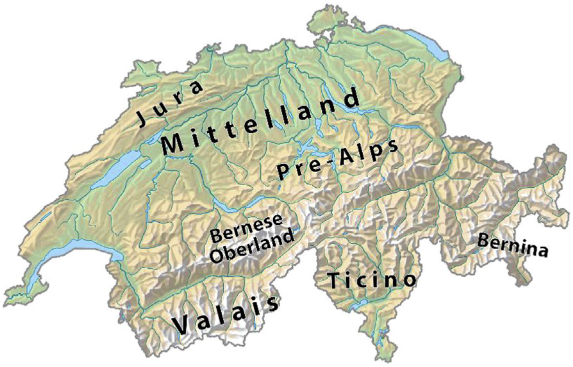

Fig. 1. Swiss topography and key regions referred to in the text.

Although it is possible to observe interesting and meaningful patterns in our term vectors by inspection, they also lend themselves well to more quantitative analysis. Table 2 shows correlation values calculated for the eight grid cells shown in Fig. 4. The matrix on the left represents correlation values within a given spatial folksonomy, such that for instance each HIKR region is compared to the other seven HIKR regions (upper right half of the table). Correlations within Text&Berg show slightly higher correlations, presumably as the articles in Text&Berg are more coherent in terms of undergoing a formal editorial process. Also, HIKR contains descriptions of a broad selection of outdoor activities, whereas in Text&Berg the major focus is clearly on mountaineering. Both within folksonomy comparisons show higher correlations for pairwise comparisons of regions that are expected to be more similar given their landscape characteristics.

The matrix on the right of Table 2 represents between spatial folksonomy correlations. As is expected, between folksonomy comparisons typically result in lower correlation values as compared to correlations within folksonomies. Importantly this matrix is not symmetric since rows and columns with the same labels (e.g. Matterhorn) represent information from our two spatial folksonomies (i.e. Text + Berg and HIKR). About 25% of all pairwise correlations are statistically significant (Table 2). Lower correlations typically result from comparisons of relatively diverse landscapes, such as for instance from comparing Finsteraarhorn to Lenzerheide, Toggenburg or Thun. By contrast comparisons of the same landscape described by a different spatial folksonomy (diagonal from upper left to lower right), in general give high correlations (and in all cases statistically significant). Table 2 indicates that the character of a landscape, captured in the texts of our two corpora, has stronger impact on the spatial folksonomies than the data source.

Fig. 5 illustrates the correlation between Text + Berg and HIKR per grid cell across the whole of Switzerland. In regions with limited data (c.f. Fig. 2), and thus sparse term vectors, correlations are generally lower, whilst in the regions forming the focus of our two corpora relatively high correlations of between 0.4 and 0.6 are typical, showing that the corpora describe landscapes in broadly similar, but not identical ways, where coverage is adequate.

Table 1
Basic properties of the spatial folksonomies.

	HIKR	Text + Berg
Total number of articles/	25,000/	10,000/
tokens analysed	9,000,000	22,000,000
Total number of toponyms identified/	150,000/	300,000/
median number of toponyms per article	4	11
Total number of cells in which natural features were found	461 (of 488)	466 (of 488)
Median number of natural features per cell	140	1042
Median number of unique natural features per cell	34 (of 93)	79 (of 94)

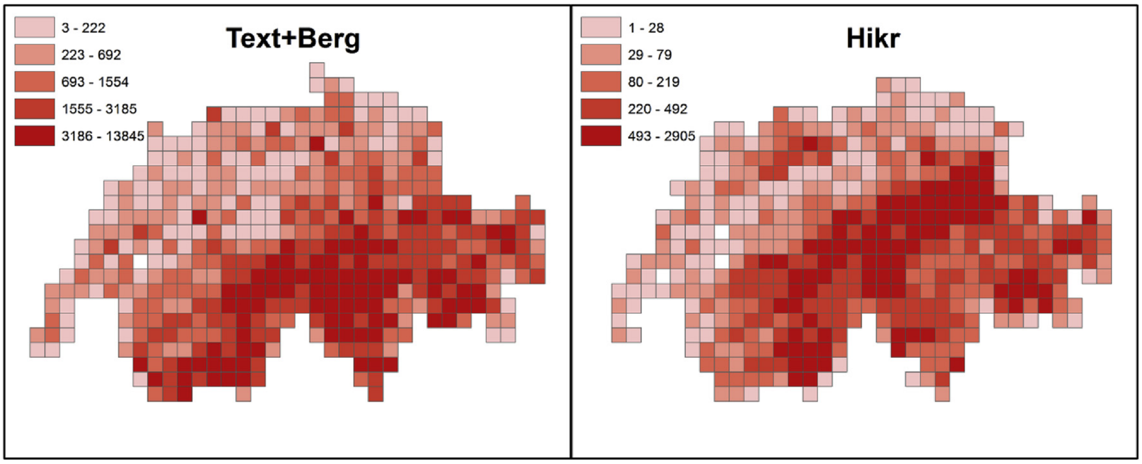

Fig. 2. Total count of natural feature terms per grid cell.

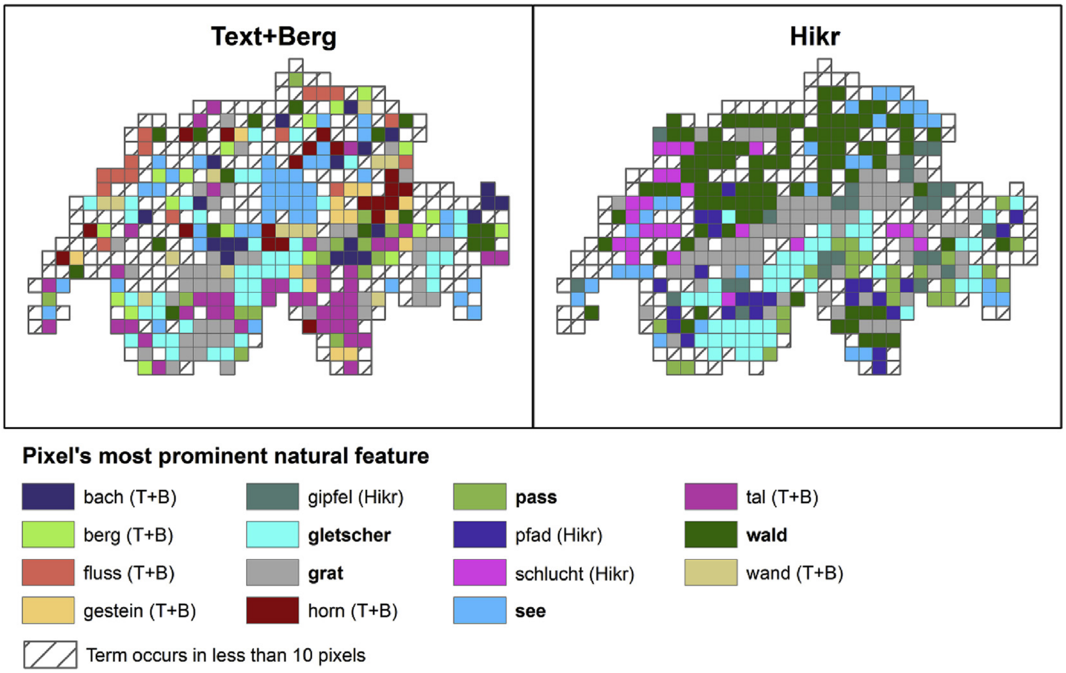

Fig. 3. Most prominent natural features per grid cell (maximum tf-idf value) for Text + Berg and HIKR.

4.2. *Spatial folksonomies and land cover classifications*

Having demonstrated that informative and useful information is stored in the spatial folksonomies, we can pose the question as to how these data compare to existing continuous classifications of, for example, land cover.

Fig. 6 shows the same information represented in Fig. 4 for four of the original eight regions, but for these four regions the 15 (where available) most frequent classes from CORINE (COR) and Arealstatistik (AS) are represented (font size represents frequency) along with the most prominent natural features from the two spatial folksonomies.

Based on Fig. 6, four observations can be made. Firstly, class labels used in CORINE and Arealstatistik are defined for the purpose of summarizing similar types of land cover/land use and for being as distinctive as possible. They are thus not meant to represent common sense landscape terms as for instance used in natural landscape and textual descriptions. Secondly, the content of the spatial folksonomies and the land cover/land use classifications show considerable overlap with for instance the classes Gletscher_Firn and Glaciers_perp_Snow prominently co-occurring with the spatial folksonomy term gletscher (glacier) in the two mountain regions *Finsteraarhorn* and *Matterhorn*. Thirdly, the land cover/land use classifications, not being restricted to exclusively using natural landscape classes, also include artificial features in the region containing the city of Thun (e.g. Discontinuous_Urban_Fabric in CORINE or Ackerland (arable land) in Arealstatistik). Finally, the land cover/land use data retrieved for the two mountain landscapes is sparse, compared to the rich descriptions contained in the spatial folksonomies, and the distribution of terms is much more skewed than is the case in our spatial folksonomies. Glacier classes have by far the highest occurrence in both regions. In particular in the Finsteraarhorn region only two out of 44 available CORINE classes occur, characterising this 10 × 10 km spatial extent as exclusively

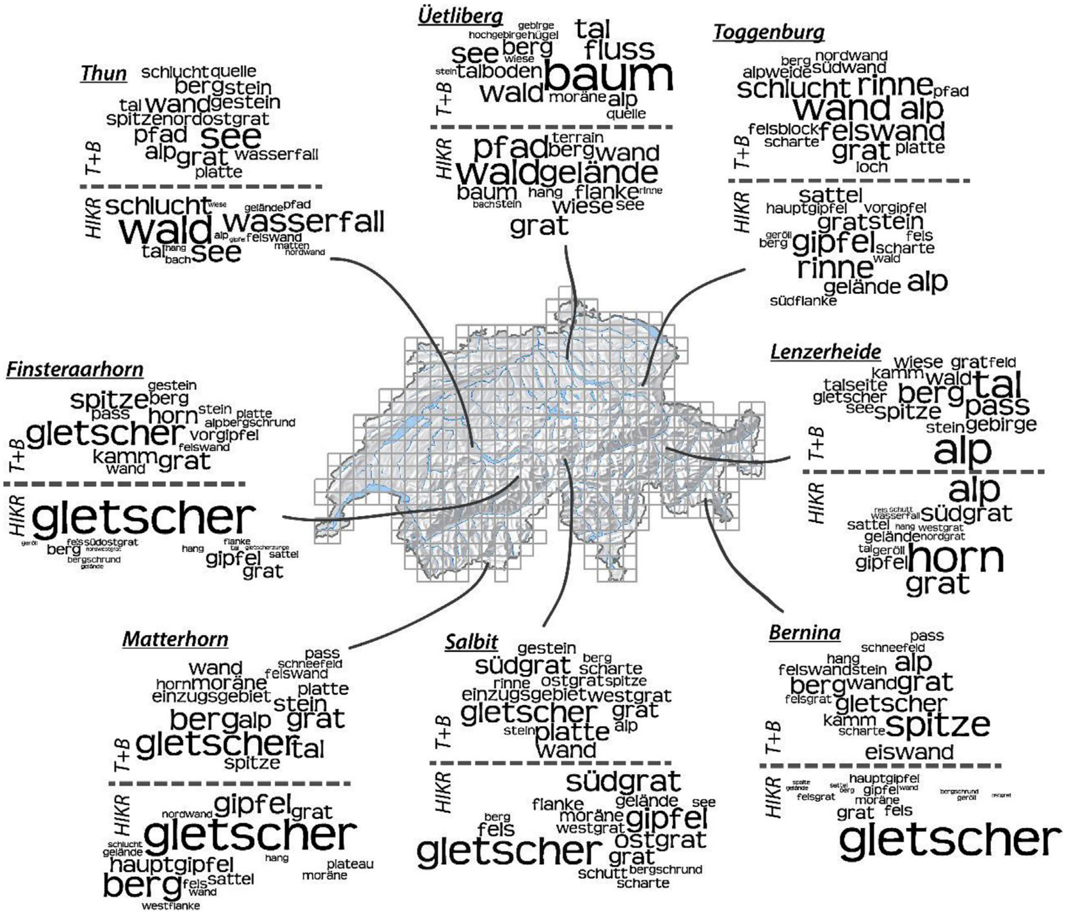
Üetliberg
gebirge hochgebirge hügel tal see berg wiese fluss stein talboden baum wald moräne alp quelle
T+B
HIKR
pfad terrain berg wand wald gelände baum hang flanke rinne bachstein wiese see grat
Toggenburg
berg nordwand alpweide südwand schlucht rinne pfad wand alp felsblock felswand scharte grat platte loch
T+B
HIKR
sattel hauptgipfel vorgipfel gratstein geröll berg gipfel fels scharte wald rinne gelände alp südflanke
Thun
schlucht quelle bergstein tal wand gestein spitze nordostgrat pfad see alp grat wasserfall platte
T+B
HIKR
schlucht wiese gelände pfad wald wasserfall alp gipfel felswand tal hang bach see matten nordwand
Finsteraarhorn
gestein spitze berg pass horn stein platte alp bergschrund gletscher vorgipfel felswand kamm wand grat
T+B
HIKR
gletscher geröll fels südostgrat berg nordwestgrat bergschrund gelände hang flanke tal gletscherzunge gipfel sattel grat
Lenzerheide
wiese grat feld kamm wald tal talseite gletscher berg see spitze pass stein gebirge alp
T+B
HIKR
alp fels schutt wasserfall südgrat sattel hang westgrat gelände nordgrat tal geröll gipfel horn grat
Matterhorn
pass schneefeld wand felswand horn moräne platte einzugsgebiet stein berg alp grat gletscher tal spitze
T+B
HIKR
nordwand gipfel grat schlucht gelände gletscher hauptgipfel hang plateau moräne berg fels sattel wand westflanke
Salbit
gestein berg südgrat scharte rinne ostgrat spitze einzugsgebiet westgrat gletscher grat stein platte alp wand
T+B
HIKR
südgrat flanke gelände see berg moräne fels westgrat gipfel gletscher ostgrat grat schutt bergschrund scharte
Bernina
pass schneefeld hang felswand stein alp berg wand grat felsgrat gletscher kamm scharte spitze eiswand
T+B
HIKR
spalte gelände sattel berg hauptgipfel gipfel wand felsgrat moräne bergschrund geröll ostgrat grat fels gletscher

Fig. 4. Spatial word clouds representing the 15 most prominent natural features separately for both spatial folksonomies and eight grid cells representing diverse mountain landscapes in Switzerland. Toponyms representative of each region are underlined at top left of each word cloud. Font size represents tf-idf value.

being covered with glacier, firn and rock.

Fig. 7 compares the spatial coverage of the two administrative land cover classifications in Switzerland and the spatial folksonomies, visualising the relative number of terms used to characterise each cell of the 10 km grid. For the spatial folksonomies, landscape terms are represented by 96 natural feature terms. The two land cover classifications use different sets of terms; Arealstatistik has a total of 72 and CORINE 44 possible classes.

Table 2
Correlation values between natural feature term vectors calculated for the eight grid cells shown in Fig. 4— *italic values* indicate statistically significant differences (Kolmogorov-Smirnov test with p-value<0.05). Left: Within HIKR and Text&Berg correlations. Right: Across HIKR and Text&Berg correlations.

HIKR - HIKR

	Finsteraarhorn	Matterhorn	Salbit	Bernina	Lenzerheide	Toggenburg	Uetliberg	Thun
Finsteraarhorn		0.45	0.46	0.46	0.37	0.28	0.28	0.24
Matterhorn	0.70		0.53	0.57	0.35	0.49	*0.22*	0.23
Salbit	0.56	0.53		0.57	0.43	0.50	*0.17*	*0.14*
Bernina	0.75	0.62	0.53		0.39	0.44	0.24	*0.21*
Lenzerheide	0.39	0.46	0.29	0.38		0.56	0.42	0.30
Toggenburg	0.42	0.52	0.64	0.53	0.59		0.37	0.32
Uetliberg	*0.19*	0.36	*0.00*	*0.07*	0.48	0.40		0.46
Thun	0.42	0.54	0.33	0.35	0.44	0.31	0.31	

T&B - T&B

HIKR T&B	Finsteraarhorn	Matterhorn	Salbit	Bernina	Lenzerheide	Toggenburg	Uetliberg	Thun
Finsteraarhorn	0.33	0.40	0.30	0.32	*0.16*	*0.16*	0.30	*0.16*
Matterhorn	0.29	0.37	0.31	0.26	*0.14*	0.32	0.26	0.33
Salbit	*0.17*	0.41	0.47	0.26	0.33	0.41	0.43	0.37
Bernina	0.23	0.39	0.34	0.48	*0.18*	*0.18*	0.35	0.24
Lenzerheide	*0.17*	*0.07*	*0.16*	*0.05*	0.25	0.28	*0.11*	0.33
Toggenburg	0.22	0.31	0.37	*0.10*	0.38	0.44	*0.24*	0.41
Uetliberg	0.26	*0.16*	0.38	*0.21*	0.26	0.41	0.38	0.37
Thun	*0.21*	0.24	0.34	*0.15*	0.34	0.40	0.36	0.33

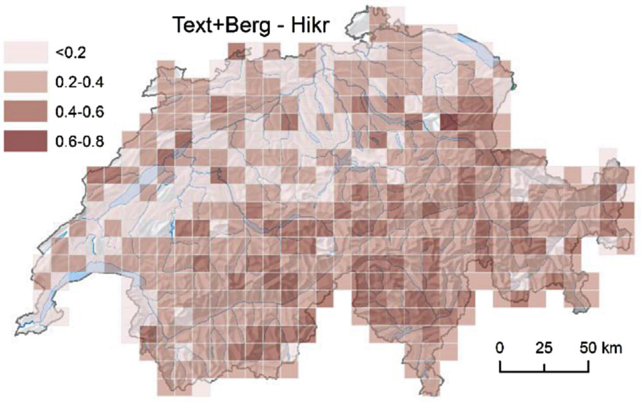

Fig. 5. Feature vector correlations between grid cells of HIKR and Text + Berg.

CORINE and Arealstatistik have similar properties, with a few important differences. Firstly, most class diversity is in more populated areas (e.g. around Uetliberg and Thun). However, while Arealstatistik has less diversity in mountain areas, this is not the case for CORINE, where the areas with least diversity are actually found on the Mittelland plains, presumably due to the relatively limited number of agricultural and other activities found here from a European perspective. Indeed, the richest classifications appear to be Arealstatistik and Text + Berg, but a note of caution is necessary here. Since Arealstatistik is a Swiss classification then all classes must, by definition, be represented, while the 96 landscape terms used in our classification were derived from Text + Berg. CORINE is a European classification, and a variety of classes are not present in Switzerland (e.g. for example any marine coastally related classes), while in our analysis of HIKR we used landscape terms derived from Text + Berg.

4.3. Temporal and contextual variation in spatial folksonomies

In a final analysis, visualised in Fig. 8, we explored how our spatial folksonomies can be stratified according to time and theme using Text + Berg and HIKR respectively, and present the most prominent natural features per grid cell (as in Fig. 3). We do this by selecting only resources associated with a particular theme in HIKR, or a particular timestamp in Text + Berg, as an initial preprocessing step prior to the generation of the folksonomies.

HIKR landscape descriptions originating from articles describing hiking show both a spatial and thematic shift, covering all of Switzerland, and commonly referring to *wald* (forest), *see* (lake) and *grat* (ridge). In articles more focussed on mountaineering, forest and lakes more or less completely disappear, to be replaced by *gletscher* (glacier) and a much more tight spatial focus centred around the Swiss Alps.

Similarly, temporal stratification of Text + Berg reveals a key change of focus to *gletscher* (glacier) post-1950. This result does not suggest changes *per se* in glaciers, but rather a change of emphasis

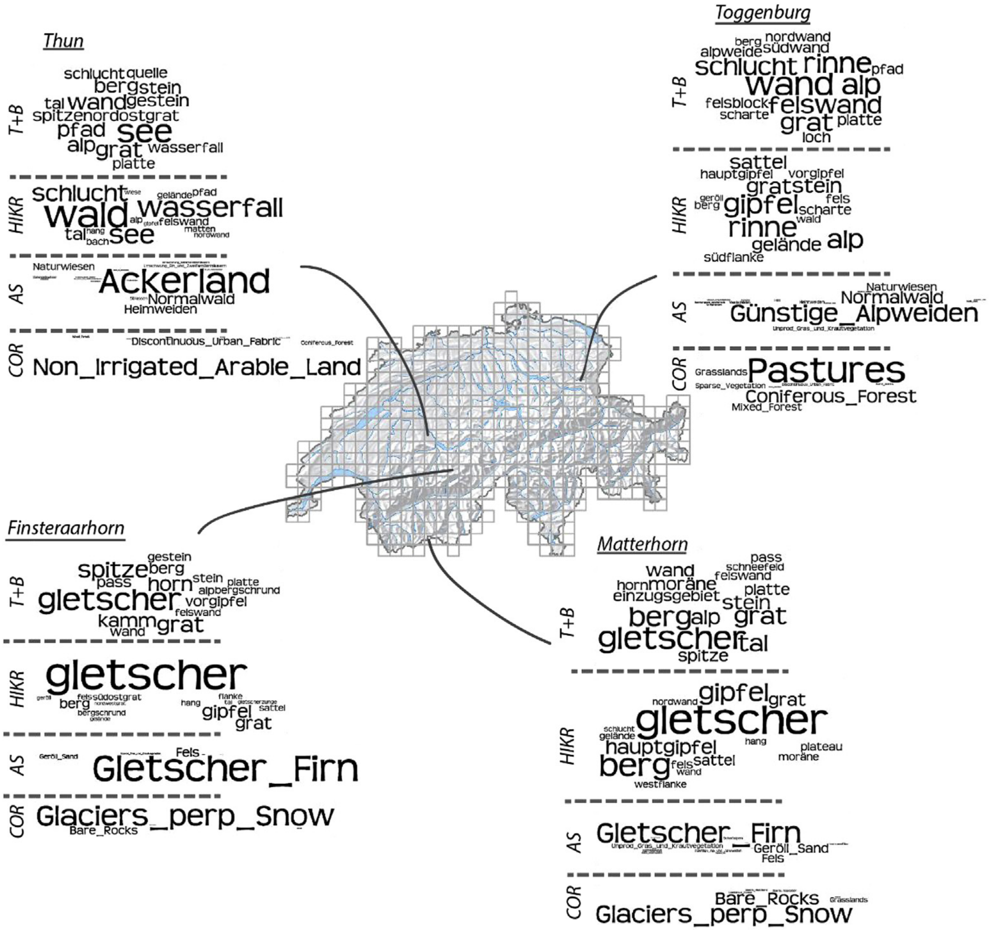

Fig. 6. Regional comparison of the 15 most frequent classes from CORINE (COR) and Arealstatistik (AS) with the 15 most prominent natural feature terms from the two spatial folksonomies.

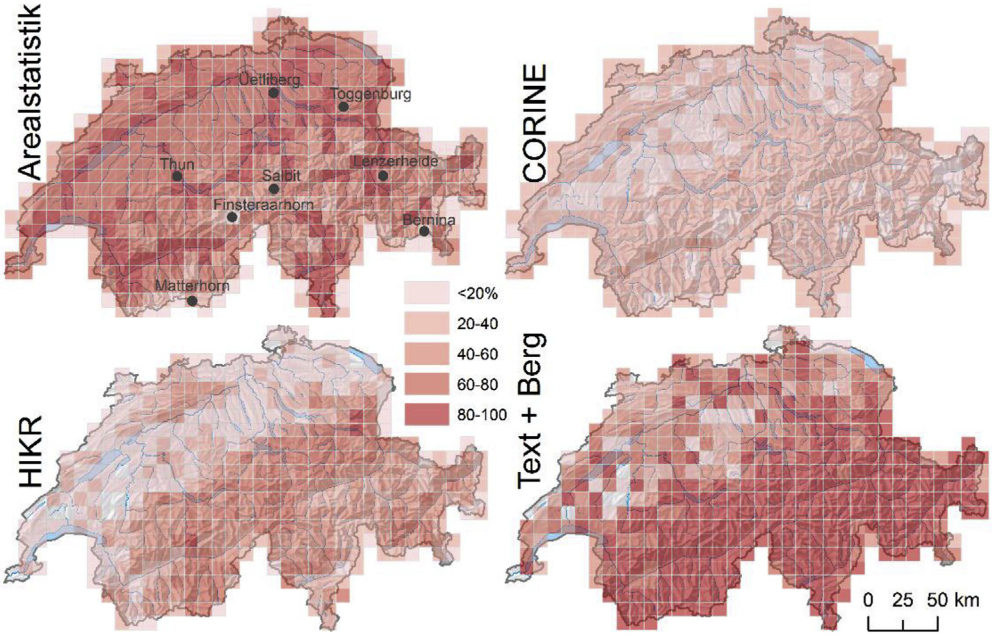

Fig. 7. Relative use of the available land cover classes and natural features, respectively, for each grid cell.

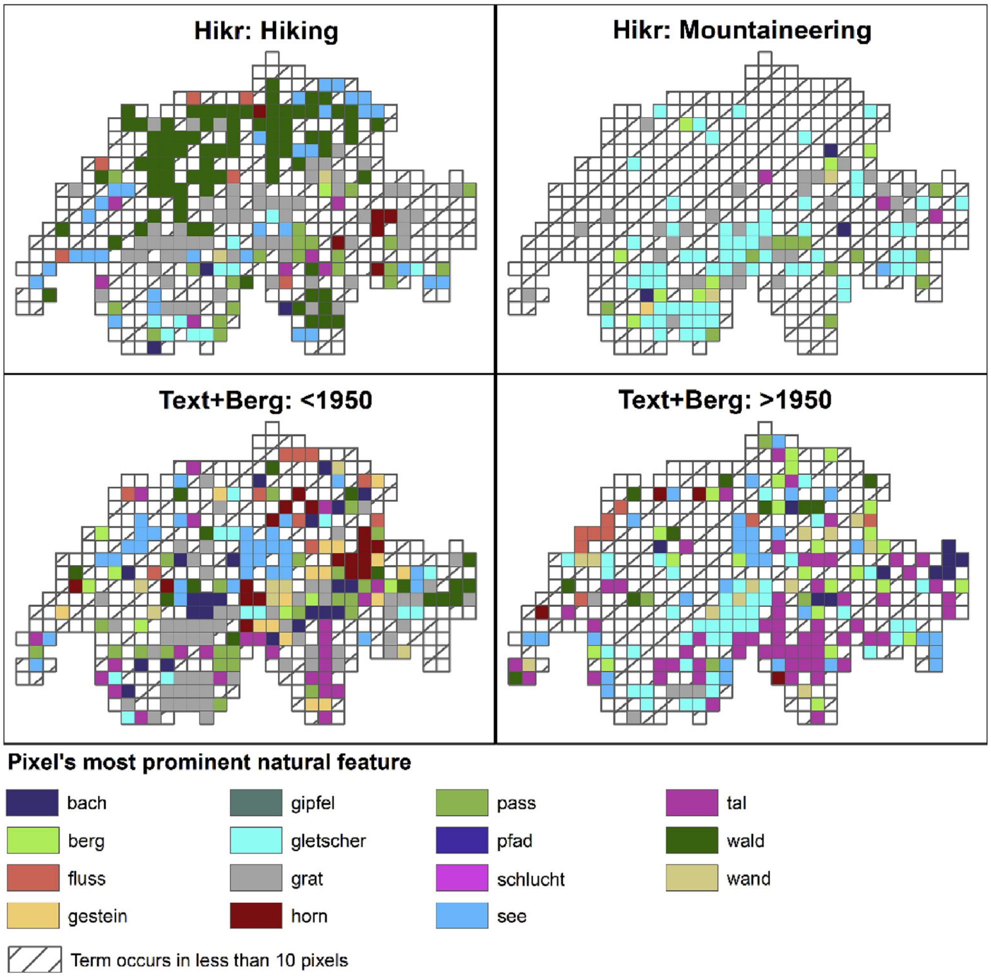

Fig. 8. Stratifications of the two spatial folksonomies, using topic (i.e. activity, HIKR) and year of publication (Text + Berg).

in our texts to make these a particularly prominent landscape feature. This fits well with observations made by, for example, Haeberli (2009), who has argued that perceptions of glaciers fundamentally changed in the course of the 20th century, reflecting the process of climate change through deglaciation, with glaciers becoming “unique demonstration objects”.

5. Discussion

In our introduction we argued that current land use, land cover and landscape classifications typically reflected a top-down process, especially with respect to the involvement of non-experts. Our approach to addressing this issue harnessed neither crowd sourcing (Fritz et al., 2009), nor the currently popular notions of UGC and volunteered geographic information *per se* (Dykes, Purves, Edwardes, & Wood, 2008; Goodchild, 2007), but rather textual sources from two contrasting written corpora. Such corpora are very rich, but also complex, and we believe they have considerable potential as sources of geographic information. In this paper, we used these corpora to produce what we call a spatial folksonomy, essentially a set of georeferenced, ranked term vectors describing natural landscape features in Switzerland. In previous work (Derungs & Purves, 2014) we described the process of extracting these term vectors for one of our corpora (Text + Berg) in some detail. Here we set out to explore how these term vectors could be used to form spatial folksonomies, and in particular the related reproducibility (RQ1), comparability (RQ2) and utility (RQ3) of such data structures.

Although the broad methodological foundations for deriving our spatial folksonomy were laid out in Derungs & Purves, (2014), a key question in this paper concerned the extent to which we could apply such methods both automatically and reproducibly on other corpora. Doing so required that we also made compromises – in our previous work we annotated a set of natural landscape features derived directly from the corpora, and used a spatially adaptive grid to position information. However, since the set of landscape features will obviously vary according to individual corpora, and the adaptive grid reflects variations in spatial densities of descriptions across Switzerland, we compromised in this paper and used the list of 96 natural landscape features identified in Text + Berg, and a regular 10 km grid.

These compromises manifest themselves in a number of ways. Firstly, Text + Berg has more natural features per grid cell (Table 1 & Fig. 2). However, the list of natural features is extensive enough such that rich, meaningful, but different, descriptions also emerge in HIKR (Figs. 3 and 4), allowing us to contrast ways in which landscapes are described. A key test is whether these descriptions are consistent. This is best illustrated in Table 2, where we found statistically significant correlations between natural term feature vectors at a grid cell level for 8 exemplar grid cells, and Fig. 5, showing relatively strong correlations between natural term feature vectors, especially in the Alps.

Another obvious limitation of our method concerns the arbitrary nature of the grid cells – a classic manifestation of the Modifiable Areal Unit Problem (MAUP: Openshaw, 1983). Approaches to dealing with MAUP include varying the choice of aggregation method for raster cells (e.g. Schmit, Rounsevell, & La Jeunesse, 2006), or using more meaningful ecological or perceptive units to parcel up terms such as catchments (Lausch & Herzog, 2002) or viewsheds (Nahuelhual, Carmona, Lozada, Jaramillo, & Aguayo, 2013). Furthermore, our characterisation has very coarse spatial resolutions (Lausch & Herzog, 2002). However, 10 km is a broadly meaningful aggregate scale at which to capture and compare landscape conceptualisations providing an overview which may help in identifying areas of landscapes with broadly similar characteristics for more detailed comparison (c.f. Brown, Raymond, & Corcoran, 2015).

If we are to use spatial folksonomies in conjunction with, or as extensions to, current landscape categorisations such as CORINE and Arealstatistik then we must ask ourselves whether these representations are in any way comparable and, if so, what added value can be gained from the spatial folksonomies. Answers to both these questions are given in Fig. 6, which allowed us to conclude that the semantic content of the two data types is of comparable nature, with the spatial folksonomies however giving a richer description of mountain regions such as Finsteraarhorn, which through the prism of current landscape categorisations are made up of glaciers and rock. The distribution of semantic coverage over the area of Switzerland is well illustrated by Fig. 7, were we find additional evidence that the spatial folksonomies are in general richer in mountainous regions, in contrast to CORINE and Arealstatistik, which tend to have most class diversity in urban and lower-lying agricultural areas.

Having established that our method allows us to compare landscapes (at an admittedly coarse scale) has important implications. It is generally becoming accepted in remote sensing that, rather than striving to achieve the perfect classification model with a single product, integrating additional geographical data is a powerful way of improving classification models (Rozenstein & Karnieli, 2011). Our approach allows us to even integrate implicit spatial data and to capture terms used to describe landscapes in textual corpuses produced by non-experts.

The final research question was concerned with the effectiveness of spatial folksonomies in enabling discussions on landscapes, land use and land cover. By focussing on broad regions, and using rich natural feature term vectors we argue firstly that it is possible to characterise, and thus discuss, landscapes in quite different ways to traditional land cover data. For instance, the classes of ridge and forest which emerge as particularly prominent members in Fig. 3 (especially for HIKR) seem to suggest classes of locations affording quite different forms of hiking in, respectively, the relatively flat Mittelland, and the more mountainous Pre-Alps area. This notion of affordances, that is to say the ways in which an environment supports activities *sensu* Gibson (1979) has parallels with much more recent debates on cultural ecosystem services. Our spatial folksonomies therefore start to provide us with access in particular to notions of important natural features terms with respect to both recreation and aesthetics, which are not easily extractable from purely biophysical data (González-Puente, Campos, McCall, & Muñoz-Rojas, 2014; Thornton, 2011). Currently the need to capture such information spatially is mostly addressed either using participatory and interview based approaches (e.g. Brown et al., 2015; Sherrouse, Clement, & Semmens, 2011) and/or deterministic modelling using Geographic Information Systems (Kuenzer & Tuan, 2013; Nahuelhual et al., 2013). The approach presented in this paper offers a novel third way. The design and population of spatial folksonomies from landscape descriptions and the direct access offered to shared conceptualisations of landscapes from text is a potential new source for ‘cultural’ information. Furthermore, where corpora have existed over long time periods, contain fine spatial granularity information, and can be stratified in other ways (for example by activities, c.f. Fig. 8), it may be possible to start to explore cultural ecosystem services across geographic and temporal scales (Carpenter et al., 2006) and to consider different ways in which a landscape is perceived. Analysing rich textual resources, created by large communities of authors offers a possible approach to better capturing the variety of ways in which landscapes are described, and thus a route to addressing the need for reflecting non-expert conceptualisations of landscapes in policy (c.f. Prieur et al., 2006).

Our approach has a number of important limitations. Firstly, and most obviously we worked with Swiss corpora, and further work would be needed to firstly identify suitable corpora, and secondly extend our methods to work with these across broader regions. Here, lack of availability of homogenous, fine granularity gazetteers (the popular Geonames does not meet this requirement) remain a challenge. Secondly, our approach to identifying natural features was manual, and based on inter-annotator agreement. Clearly the list itself is to some extent corpus dependent, and it would be interesting to, for example, crowdsource such natural features. Furthermore, since landscapes are typically defined as coupled human and natural systems, it would make sense to extend the list beyond purely natural features. Thirdly, although our corpora allowed us to stratify both temporally and thematically, we observed that we were rapidly confronted with a problem of data scarcity despite the initially large and rich corpora. Fourthly, we note that the link between natural features and locations is simplistic – for example we do not differentiate between being in and seeing a landscape from afar. Current advances in Geographic Information Retrieval have considerable promise here (Moncla, Renteria-Agualimpia, Nogueras-Iso, & Gaio, 2014). Finally, we do not consider the potential impacts of individual users making large contributions to our datasets. Though this is unlikely to be an issue for Text + Berg, simply through its very long temporal span, we will explore the impact of individual authors in HIKR in more detail in future work.

6. Conclusion and outlook

We opened this paper arguing for a need to develop bottom-up approaches to describing landscapes, land cover and land use. Such approaches have the potential to better capture local variation in the ways in which landscapes are described, and thus also potentially better meet local needs, while dealing with the challenge of ontological mismatches between seemingly transparent terms such as forest (Comber et al., 2005).

Our approach to meeting this challenge was to develop what we termed spatial folksonomies for Switzerland using two, thematically similar, but quite different textual corpora. We argued that such corpora contain very rich information, in our case allowing us to build spatial folksonomies containing natural feature terms at a resolution of 10 km. Our approach is a novel one, using full text corpora as a starting point to generate rich, spatially referenced, landscape descriptions. Although we have only scratched the surface of the potential of exploring such methods, we believe our approach has a number of important implications which are demonstrated in this paper.

Firstly, the state of the art in methods from Geographic Information Retrieval is now such that, subject to availability of suitable corpora and methods for identifying relevant terms, it is possible to generate meaningful spatial folksonomies. Using diverse, rich textual corpora we captured, at a relatively coarse granularity, variation in descriptions of (mountain) landscapes through natural features in Switzerland. Although the nature of the terms describing regions vary according to individual corpora, descriptions created using the same lists of natural features correlate in space. Thus, our approach can be used to identify similar regions using standard methods for comparing documents.

Secondly, our spatial folksonomies contain rich descriptions of grid cells which relate well to more informal ways of describing landscapes. As such, our spatial folksonomy provides a useful way of generating spatial queries using folk landscape terms (such as ridge, summit or forest) and ranking documents according to the prevalence of such terms in a particular region. Furthermore, our terms complement traditional land cover and land use datasets, particularly in so-called unproductive areas where land cover and land use classifications are often relatively sparse. We see a particularly important usage of our work in providing ways of exploring the appropriateness of current land cover and land use classifications at a local level, and in providing ways of linking rich text descriptions to existing ways of classifying land cover and land use.

Thirdly, we demonstrated by stratifying our corpora thematically (through different activities) and temporally, variation in the ways landscapes were described. By extending our work to larger regions, such stratification can become particularly interesting, for example in exploring cultural ecosystem services and their relationship to particular activities, or ways in which perception of landscapes has varied over time. Such research should extend beyond lists of nouns related to natural features to include other terms related to cultural uses of landscapes and their characterisation.

Acknowledgements

We would like to thank the University Research Priority Program *Language and Space* of the University of Zurich and the project *FolkOnt* funded by the Swiss National Science Foundation under contract 200021-126659 for supporting this work. Finally, we are grateful to the computational linguistics group of Prof. Martin Volk at the University of Zurich for allowing access to the digitized and preprocessed Text + Berg corpus.

References

Ahern, S., Naaman, M., Nair, R., & Yang, J. H.-I. (2007). World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries* (pp. 1–10).

Brown, G., Raymond, C. M., & Corcoran, J. (2015). Mapping and measuring place attachment. *Applied Geography, 57*, 42–53.

Cantador, I., Bellogín, A., & Vallet, D. (2010). Content-based recommendation in social tagging systems. In *Proceedings of the fourth ACM conference on recommender systems* (pp. 237–240).

Carpenter, S. R., De Fries, R., Dietz, T., Mooney, H. A., Polasky, S., Reid, W. V., et al. (2006). Millennium ecosystem assessment: Research needs. *Science, 314*(5797), 257–258.

Comber, A., Fisher, P., & Wadsworth, R. (2005). What is land cover? *Environment and Planning B: Planning and Design, 32*(2), 199–209.

Derungs, C., & Purves, R. S. (2014). From text to landscape: Locating, identifying and mapping the use of landscape features in a Swiss Alpine corpus. *International Journal of Geographical Information Science, 28*(6), 1272–1293.

Dykes, J., Purves, R. S., Edwardes, A., & Wood, J. (2008). Exploring volunteered geographic information to describe Place: Visualization of the 'Geograph British Isles' collection. In *Proceedings of the GIS research UK 16th annual conference GISRUK* (pp. 256–267).

EEA-ETC. (1994). *CORINE land cover*.

Feddema, J. J., Oleson, K. W., Bonan, G. B., Mearns, L. O., Buja, L. E., Meehl, G. A., et al. (2005). The importance of land-cover change in simulating future climates. *Science, 310*(5754), 1674–1678.

Feranec, J., Jaffrain, G., Soukup, T., & Hazeu, G. (2010). Determining changes and flows in European landscapes 1990–2000 using CORINE land cover data. *Applied Geography, 30*(1), 19–35.

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., et al. (2009). Geo-Wiki. org: The use of crowdsourcing to improve global land cover. *Remote Sensing, 1*(3), 345–354.

Gibbs, H. K., & Salmon, J. M. (2015). Mapping the world's degraded lands. *Applied Geography, 57*, 12–21.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin Company.

González-Puente, M., Campos, M., McCall, M. K., & Muñoz-Rojas, J. (2014). Places beyond maps; integrating spatial map analysis and perception studies to unravel landscape change in a Mediterranean mountain area (NE Spain). *Applied Geography, 52*, 182–190.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211–221.

Gruber, T. R. (2007). Folksonomy of ontology: A mash-up of apples and oranges. *International Journal on Semantic Web & Information Systems, 3*(2), 1–11.

Haeberli, W. (2009). *Gletscherschwund – Verlust eines Mythos?. 66* pp. 221–228) Mitteilungen der Naturforschenden Gesellschaft in Bern.

Hollenstein, L., & Purves, R. S. (2010). Exploring place through user-generated content: Using flickr to describe city cores. *Journal of Spatial Information Science, 1*(1), 21–48.

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. *The Semantic Web: Research and Applications*, 411–426.

Hotz, M.-C., & Weibel, F. (2005). *Arealstatistik Schweiz: Zahlen-Fakten-Analysen*. Neuchatel.

Kuenzer, C., & Tuan, V. Q. (2013). Assessing the ecosystem services value of can gio mangrove biosphere reserve: Combining earth-observation-and household-survey-based analyses. *Applied Geography, 45*, 167–184.

Lausch, A., & Herzog, F. (2002). Applicability of landscape metrics for the monitoring of landscape change: Issues of scale, resolution and interpretability. *Ecological Indicators, 2*(1), 3–15.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014), 176–182.

Moncla, L., Renteria-Agualimpia, W., Nogueras-Iso, J., & Gaio, M. (2014). Geocoding for texts with fine-grain toponyms: An experiment on a geoparsed hiking descriptions corpus. In *Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 183–192).

Nahuelhual, L., Carmona, A., Lozada, P., Jaramillo, A., & Aguayo, M. (2013). Mapping recreation and ecotourism as a cultural ecosystem service: An application at the local level in Southern Chile. *Applied Geography, 40*, 71–82.

Openshaw, S. (1983). *The modifiable areal unit problem*. Norwich.

Perger, C., Fritz, S., See, L., Schill, C., Van Der Velde, M., McCallum, I., et al. (2012). A campaign to collect volunteered geographic information on land cover and human impact. *GI_Forum*, 83–91.

Price, B., Kienast, F., Seidl, I., Ginzler, C., Verburg, P. H., & Bolliger, J. (2015). Future landscapes of Switzerland: Risk areas for urbanisation and land abandonment. *Applied Geography, 57*, 32–41.

Prieur, M., Luginbühl, Y., Zoido Naranjo, F., De Montmollin, B., Pedroli, B., Van Mansvelt, J. D., et al. (2006). *Landscape and sustainable development-challenges of the European landscape convention*.

Purves, R. S., & Jones, C. (2006). Geographic information retrieval (GIR). *Computers, Environment and Urban Systems, 30*(4), 375–377.

Rozenstein, O., & Karnieli, A. (2011). Comparison of methods for land-use classification incorporating remote sensing and GIS inputs. *Applied Geography, 31*(2), 533–544.

Schmit, C., Rounsevell, M. D. A., & La Jeunesse, I. (2006). The limitations of spatial land use data in environmental analysis. *Environmental Science & Policy, 9*(2), 174–188.

Sennrich, R., Schneider, G., Volk, M., & Warin, M. (2009). A new hybrid dependency parser for German. In *Proceedings of GSCL-conference. Potsdam* (pp. 115–124).

Sherrouse, B. C., Clement, J. M., & Semmens, D. J. (2011). A GIS application for assessing, mapping, and quantifying the social values of ecosystem services. *Applied Geography, 31*(2), 748–760.

Spinsanti, L., & Ostermann, F. (2013). Automated geographic context analysis for volunteered information. *Applied Geography, 43*, 36–44.

Steinmeier, C. (2013). *CORINE land cover 2000/2006. Birmensdorf, Switzerland*.

Thornton, T. F. (2011). Language and landscape among the Tlingit. In D. M. Mark, A. G. Turk, N. Burenhult, & D. Stea (Eds.), *Landscape in language: Transdisciplinary perspectives* (pp. 275–289). Philadelphia: John benjamins Publishing Company.

Volk, M., Bubenhofer, N., Althaus, A., & Bangerter, M. (2010). Classifying named entities in an alpine heritage corpus. *Künstliche Intelligenz, 4*, 40–43.

Winget, M. (2006). User-defined classification on the online photo sharing site Flickr … Or, how I learned to stop worrying and love the million typing monkeys. *Advances in Classification Research Online, 17*(1), 1–16.