An exploratory analysis of usability of Flickr tags for land use/land cover attribution

Yingwei Yan, Michael Schultz and Alexander Zipf

Institute of Geography, Heidelberg University, Heidelberg, Germany

ABSTRACT
This study explored the land use/land cover (LULC) separability by the machine-generated and user-generated Flickr photo tags (i.e. the auto-tags and the user-tags, respectively), based on an authoritative LULC dataset for San Diego County in the United States. Ten types of LULCs were derived from the authoritative dataset. It was observed that certain types of the reclassified LULCs had abundant tags (e.g. the parks) or a high tag density (e.g. the commercial lands), compared with the less populated ones (e.g. the agricultural lands). Certain highly weighted terms of the tags derived based on a term frequency–inverse document frequency weighting scheme were helpful for identifying specific types of the LULCs, especially for the commercial recreation lands (e.g. the zoos). However, given the 10 sets of tags retrieved from the corresponding 10 types of LULCs, one set of tags (all the tags located at one specific type of the LULCs) could not fully delineate the corresponding LULC due to semantic overlaps, according to a latent semantic analysis.

1. Introduction
Human species reshapes 50% of earth’s surface and continues to accelerate land transformations (Hooke, Martín-Duque, and Pedraza 2012; Waters et al. 2016). Planetary bookkeeping of land changes for land management has become a prerequisite for permanent human existence (Woodcock and Ozdogan 2004). Specifically, land use/land cover (LULC) is a relevant variable to understand current issues such as climate change, urban management, and natural resources (Schultz et al. 2017). An abundance of LULC data has thus been available, for instance, the American National Land Cover Database, the Chinese LULC Change Database, DFD Land Use and Land Cover Product for Germany, and the European Coordination of Information on the Environment Land Cover Product (Bossard, Feranec, and Otahel 2000; Cihlar and Jansen 2001; Homer, Fry, and Barnes 2012; Liu et al. 2002; Mack et al. 2017); as well as global products (Arino et al. 2012; Fritz et al. 2003; McIver and Friedl 2002) and a new generation of 30 m land cover maps or layers (Schultz et al. 2017).

Alongside these traditional data products, volunteered geographical information increasingly has been used for LULC purposes. For instance, the conversion of OpenStreetMap (OSM) data into LULC features (Fonte et al. 2016). The combination of the above-mentioned products and OSM-derived products were created (Fonte et al. 2017) and data fusion was performed where remote sensing data was fused with OSM LULCs for filling eventual data gaps (Schultz et al. 2017). Several challenges exist in voluntarily contributed data. Incorrect thematic information (tagging), overlapping features, and spatial and temporal inhomogeneities are significant shortcomings affecting OSM credibility (Bégin, Devillers, and Roche 2013; Fonte et al. 2016; Jokar Arsanjani et al. 2015; Jokar Arsanjani and Vaz 2015). Data completeness is another challenge, which was found to depend on contributor activities (Foody et al. 2013; Neis and Zipf 2012). Multiple tags provided by different users were found to compromise legend harmonization which was necessary for comparison with other LULC products (Dorn, Törnros, and Zipf 2015; Estima and Painho 2013a; Jokar Arsanjani and Vaz 2015).

Alternatively, abundant data of images was created, shared online by the general public (See et al. 2016) and was used for applications pertinent to human activities (Antoniou, Morley, and Haklay 2010; Lee, Cai, and Lee 2014). Particularly, Flickr data was explored often. For instance, it was explored to assess the quality of LULC data (Estima and Painho 2013b); its usage in artificial neural networks for LULC attribution was tested (Zhu and Newsam 2015); its use for LULC attribution was compared with other data streams (Estima, Fonte, and Painho 2014); most recently, Flickr images were used to guide supervised classification of remote sensing data for deriving LULCs (Sithi et al. 2016).

Nevertheless, the existing studies focused on the contents of the images themselves, aiming for...
content decomposition in relation to other sources to attribute lands. Inspection and image processing of Flickr data is unfeasible for large area applications due to processing constrains. On the contrary, Flickr tags are relatively cheap to harvest and process (Yan et al. 2017). Currently it remains unexplored if the composition of Flickr tags can be used to determine LULCs reliably. In this study, an official LULC reference dataset for San Diego County in the United States was therefore used to explore the LULC separability by the machine-generated and user-generated Flickr photo tags (i.e. the auto-tags and the user-tags).

The remainder of this article is organized as follows. Section 2 presents the materials and methods of this research, followed by the corresponding results in Section 3. Section 4 discusses and concludes this work.

2. Materials and methods

2.1. Study area and the official LULC data

Considering the diversity of the LULCs of San Diego County in the United States, this county was selected as a case area for this study. An official LULC dataset covering the entire county (the date of data: 1 January 2016) was obtained from SanGIS/SANDAG GIS Data Warehouse (http://rdw.sandag.org/). The LULC dataset was consisted of the footprints of individual land parcels (e.g. a shopping mall) being classified into 105 types of LULCs. The LULC codes and definitions can be found in the online Supplementary Material. The 105 LULCs were reclassified into 10 types, including residential land (R), industrial land (I), transportation land (T), commercial land (C), public service, hospital, and school land (PHS), commercial recreation land (CR), park (P), agricultural land (A), water body (W), and others (military use land, planned land, and area under construction, denoted as O) (Table 1 and Figure 1).

2.2. Flickr data

Following the access policies and regulations of Flickr API, a PHP-based tool was developed to collect Flickr photos and the metadata. The tool retrieved Flickr data by scanning the study area using a 0.5 degrees by 0.5 degrees moving window, starting from the upper left corner. Since the Flickr API allowed for accessing a maximum of 4000 photos in a single API query execution, a window was subdivided into four equal-

| R  | I  | T  | C  | PHS | CR  | P  | A  | W  | O  |
|----|----|----|----|-----|-----|----|----|----|----|
| 1000 | 2001 | 4101 | 1501 | 6003 | 7201 | 7601 | 8001 | 9202 | 6701 |
| 1090 | 2101 | 4102 | 1502 | 6101 | 7202 | 7603 | 8002 | 9200 | 6702 |
| 1110 | 2103 | 4103 | 1503 | 6102 | 7203 | 7604 | 8003 | 9201 | 6703 |
| 1120 | 2104 | 4104 | 5001 | 6103 | 7204 | 7605 | 9101 |
| 1190 | 2105 | 4111 | 5002 | 6104 | 7205 | 7606 | 9501 |
| 1200 | 2201 | 4112 | 5003 | 6105 | 7206 | 7607 | 9502 |
| 1230 | 2301 | 4113 | 5004 | 6106 | 7207 | 7608 | 9503 |
| 1290 | 4114 | 5005 | 6109 | 7208 | 9504 |
| 1300 | 4115 | 5006 | 6501 | 7209 | 9505 |
| 1401 | 4116 | 5007 | 6502 | 7210 | 9506 |
| 1402 | 4117 | 5008 | 6503 | 7211 | 9507 |
| 1403 | 4118 | 5009 | 6801 | 9600 |
| 1404 | 4119 | 6001 | 6802 | 9700 |
| 1409 | 4120 | 6002 | 6803 |

Table 1. The 10 types of reclassified LULCs and the codes of the 105 types of the original LULCs.

Figure 1. Study area: San Diego County. Background map is OpenStreetMap, on top of which is the 10 types of reclassified LULCs.
size sub-windows (i.e. a quadtree division) in case more than 4000 photos were contained within that window. This subdivision was recursively performed until no API query returned more than 4000 photos. The Flickr data contributed during the 3 years (from 1 January 2013 to 31 December 2015) prior to the publication of the LULC dataset were collected. It was assumed that there was no or minor LULC changes during the 3 years. Focusing on the 3 years rather than 1 or 2 years also ensured the sufficiency of the Flickr data for our analysis. This resulted in 80,118 geo-tagged photos across the study area. The Flickr API returned two kinds of geolocation: (1) user-provided location; and (2) device-generated GPS location showing where a photo was taken. The former was ignored, as a user may manipulate the location of a photo either before or after the photo submission, making the location different from the true location where the photo was taken. The latter was therefore used, which was retrievable using the Flickr API, alongside the flickr.photos.getExif method.

Flickr’s auto-tagging system uses a list of predefined standard tags to automatically describe the contents of a photo through image recognition techniques. Flickr users can also tag a photo manually without following any predefined tagging standard. From the 80,118 geo-tagged photos (some photos had only auto-tags, some had only user-tags, and some had both), 61,949 photos that had auto-tags, and 55,980 photos that had user-tags were obtained. Within this subset of photos, it was noticeable of some extremely active users who contributed a big number of photos. In Figure 2, they are represented by the dots distributed toward the right end of the x-axis. These cases cause biases toward the geographic distribution of these users (Yan et al. 2017) and thus should be removed. To minimize such biases, the arithmetic mean of all the available values on the x-axis that correspond to the black dots on the curves (i.e. all the available values for the number of photo contributions per user) was calculated. The arithmetic mean for the auto-tags was 115.17, and that for the user-tags was 118.58. In both cases, all the users who contributed photos more than the mean were removed. This resulted in 49,363 photos with 163,628 auto-tags from 10,741 users, and 43,535 photos with 294,089 user-tags from 8994 users (Table 2).

2.3. Tag quantities

First, the associations between the tag quantities and the 10 types LULCs were investigated. The number of tags located at each type of the LULCs was counted and visualized. The number of tags per square kilometer (i.e. the tag density) was also calculated and visualized for each type of the LULCs.

In addition, the semantic characteristics of the tags associated with the different types of LULCs should be investigated. If Flickr tags retrieved from different LULCs can support LULC separation, the different sets of Flickr tags from the different LULCs should be different from each other significantly in terms of their semantic contents. From all the auto-tags, 535 terms (i.e. tag contents such as river and ship) were identified. For the user-tags, it was less straightforward. Since the users do not need to follow any predefined tagging standard or use any predefined tag, the user-tags are more heterogeneous and involve noises. Therefore, a preprocessing of the user-tags was needed. The tags with only digits, the tags combined with digits and words, and the tags with misspelled words were removed. This resulted in 7165 terms (much more than the terms from the auto-tags).

For each of the auto-tag terms and the user-tag terms, the number of times it appeared on each of the 10 types of LULCs (the term frequency) was calculated, resulting in a 535-row by 10-column term frequency matrix for the auto-tags and a 7165-row by 10-column term frequency matrix for the user-tags. Then, the pairwise Pearson correlation coefficients were calculated with the term frequencies (the correlation between each pair of the term frequency columns), which

![Figure 2. User-contribution distributions of the photos with auto-tags (a) and user-tags (b).](image-url)
enabled a preliminarily examination as to what extent the different sets of terms from the different types of LULCs were uncorrelated.

### 2.4. Term weighting and latent semantic analysis

To ensure a robust semantic analysis, a term frequency-inverse document frequency (TF-IDF) weighting scheme (Sehra, Singh, and Rai 2017) was applied to the terms. This resulted in a term weight-LULC matrix in which the rare terms were promoted, and the common terms were discounted based on term frequencies. The weighting method is expressed as Equation (1):

$$W_{ij} = \frac{n_{ij}}{N_j} \times \log_2(n_{LULC}/df_i)$$  \hspace{1cm} (1)

where $W_{ij}$ denotes the TF-IDF weight obtained, $n_{ij}$ denotes the number of times term $i$ appears on LULC $j$, $N_j$ denotes the total number of terms associated with LULC $j$, $n_{LULC}$ denotes the number of LULC types involved (10 in this study), and $df_i$ denotes the number of LULCs where term $i$ appears on. The first half of the equation is a local component regarding term frequency, and the second half is a global component that describes the importance of term $i$.

In this way, a 535-row by 10-column term weight matrix for the auto-tags and a 7165-row by 10-column term weight matrix for the user-tags were obtained. With the term weight matrices, first, the terms were sorted according to their weights corresponding to each type of the LULCs, then the top 20 weighted terms were extracted to investigate to what extent they were able to delineate the corresponding LULC.

Thereafter, a latent semantic analysis (LSA) was performed to further examine to what extent each set of the weighted terms (i.e. each column of the term weight matrix) was able to describe a unique type of the LULCs. LSA is a natural language processing technique which automatically organizes, understands, and summarizes a textual dataset (Sehra, Singh, and Rai 2017). The foundation of LSA is the singular value decomposition (SVD) which can create a low dimensional space of a term weight matrix while preserving the meanings of the original matrix if there are semantic similarities among the multiple sets of weighted terms (Li, Goodchild, and Raskin 2014); each dimension represents a specific semantic direction or topic.

A term weight matrix was denoted as $X$, which is subjected to the SVD to decompose it to a term-to-dimension matrix $T$, a LULC-to-dimension matrix $L$, and a diagonal matrix $S$ with singular values representing dimension loadings (i.e. strengths) appearing in descending order, expressed as Equation (2).

$$X = TSL$$  \hspace{1cm} (2)

The multiplication of $L$ and $S$ gives the factor loadings of the 10 sets of weighted terms on each dimension. By retaining the largest $k$-dimensional singular values in the matrix $S$ and setting the remaining singular values to zero, and combining the three matrices by matrix multiplication, $X$ can be represented in a reduced LSA space by $\tilde{X}$. The maximum number of singular values (i.e. the maximum number of dimensions) generated in this way is equal to the number of the LULC types (i.e. 10).

In this study, if each set of the weighted terms could describe a unique type of the LULCs, then one set would be highly different from any other set in terms of their semantics. Thereby, the loadings of the 10 dimensions would not be highly different; each dimension would represent one type of the LULCs. Each set of the weighted terms would be highly loaded on one and only one dimension, characterizing the corresponding type of the LULCs. Otherwise, the 10 types of LULCs could not be effectively differentiated and delineated by the 10 sets of weighted terms.

### 2.5. Semantic similarity

The last step was to quantify the semantic similarities among the 10 sets of weighted terms based on cosine similarity (Zhu et al. 2011), using the loadings of the 10 sets of weighted terms on each dimension (i.e. the result yielded by the multiplication of $L$ and $S$). Cosine similarity measures the cosine of the angle between two vectors (in this study, the loading vectors). For vectors with nonnegative elements, the resulting value is bounded between zero and one, where one indicates a perfect match. Mathematically, the similarity between two vectors $\hat{a}$ (in this case, the loadings of a set of weighted terms on one dimension) and $\hat{b}$ (in this case, the loadings of another set of weighted terms on another dimension) can be expressed as:

---

**Table 2. Descriptive statistics yielded from the photo collection and processing.**

| Tag      | Original user quantity | Final user quantity | Original photo quantity | Final photo quantity | Tag quantity |
|----------|------------------------|---------------------|-------------------------|---------------------|--------------|
| Auto-tag | 10,792                 | 10,741              | 61,949                  | 49,363              | 163,628      |
| User-tag | 9043                   | 8994                | 49,434                  | 43,535              | 294,089      |
\[
\text{sim}(\vec{a}, \vec{b}) = \frac{\langle \vec{a}, \vec{b} \rangle}{\|\vec{a}\| \|\vec{b}\|}
\]  

(3)

where \( \langle \cdot, \cdot \rangle \) denotes the inner product of the two vectors, \( \vec{a} \) indicates the \( L^2 \)-norm of vector \( \vec{a} \), and \( \vec{b} \) indicates the \( L^2 \)-norm of vector \( \vec{b} \).

3. Results

3.1. Quantitative patterns

The associations between the tag quantities and the LULCs are shown in Figure 3. The tag quantities vary across the 10 types of LULCs. Regarding the absolute number of tags for both the auto-tags and the user-tags, the largest tag quantity appears on the parks (P), followed by the transportation lands (T), while the smallest appears on the agricultural lands (A) (Figure 3(a)).

Regarding the number of tags per square kilometer (the tag density) for both the auto-tags and the user-tags, the largest tag quantity appears on the commercial lands (C), followed by the commercial recreation lands (CR), and the agricultural lands (A) also are associated with a very small tag density (Figure 3(b)).

The pairwise Pearson correlation values calculated with the term frequencies associated with the 10 types of LULCs are shown in Figure 4. The upper right half of the figure shows the values for the auto-tags and the lower left half shows the values for the user-tags. In general, the correlation values are very high (the lowest value is 0.757), especially for the auto-tags, meaning that the terms found on the different types of LULCs share high commonalities. This confirms the importance of the TF-IDF weighting step to discount the common terms prior to investigating the LULC separability based on the small portions of less correlated terms.

![Figure 3](image1.png)

**Figure 3.** The number of tags (a) and the number of tags per square kilometer (b) associated with each of the 10 types of LULCs.

![Figure 4](image2.png)

**Figure 4.** Pairwise Pearson correlation values calculated with the term frequencies associated with the 10 types of LULCs, which are classified using natural breaks (Jenks). The upper right half is for the auto-tags and the lower left half is for the user-tags.
3.2. Semantic patterns

Subsequently, the top 20 TF-IDF weighted auto-tag terms and user-tag terms are shown in Tables 3 and 4. The semantic relevance of the terms to the respective types of LULCs was investigated. Among each set of the top 20 terms, some of them (the bold terms) seemed to be more explicitly relevant, according to the descriptions of the LULCs (see the online Supplementary Material). Taking the auto-tags as an example, bookshelf can be found in libraries and schools which are classified into public service, hospital, and school land (PHS). For the commercial recreation lands (CR), dolphin can be found in the SeaWorld San Diego, while flamingo, bear, and tiger are all relevant to the San Diego Zoo. Both places are commercial recreation areas. In addition, the orchard and flower fields are classified into agricultural lands (A), and both orchid and tulip are relevant terms. Moreover, ship, diver, rowing, and crew boat are all explicitly relevant to water body (W).

Figure 5 displays the singular values representing the dimension loadings. For both the auto-tags and the user-tags, the 10 dimension loadings are highly different. Particularly for the auto-tags, dimension one is obviously loaded much higher than the other dimensions (Figure 5(a)). For the auto-tags and the user-tags, the loadings of the 10 sets of weighted terms across the 10 dimensions are shown in Tables 5 and 6 and are illustrated in Figure 6. Generally, one set of the weighted terms is not highly loaded on one and only one dimension. For the auto-tags, they are largely loaded on the first several dimensions, especially on the first dimension, according to the loading peaks (Table 5 and Figure 6(a)). Regarding the user-tags, the peaks are more spread out, while certain sets of the weighted terms have multiple dimensional peaks with very approximate loading values (Table 6 and Figure 6(b)). For example, the set of the weighted user-tag terms from the industrial lands (I) is loaded similarly on dimensions two and three (0.01,326 versus 0.01,318), and that from the water body (W) is loaded similarly on dimensions four and five (0.01,133 versus 0.01,143).

Moreover, Figure 7 shows the semantic similarities among the 10 sets of weighted terms based on cosine similarity. The upper right half of the figure shows the values for the auto-tags and the lower left half shows the values for the user-tags. The semantic similarities for the auto-tags are generally higher than those for the user-tags. For certain sets of the weighted auto-tag terms, the pairwise semantic similarities can be as high as 0.632, suggesting high overlapping semantics. For certain sets of the weighted user-tag terms, the semantic overlaps can be as high as 0.329.

4. Discussion and conclusions

4.1. Research findings

This study explored if and how Flickr tags can be used to separate LULCs reliably. The analysis was conducted both quantitatively and semantically. Quantitatively, it was observed that some types of the LULCs had a greater number of tags such as the parks (P) and the commercial recreation lands (CR) or a high tag density such as the commercial (C) and the commercial recreation lands (CR) (Figure 3). However, certain types of the LULCs had only a small number of tags or a very low tag density such as the agricultural lands (A) (Figure 3). This provided some useful information for the LULC separation.

The Pearson correlation values were very high, while they were not fully correlated (Figure 4). Therefore, the TF-IDF weighting was used to promote the rare tags and discount the common tags. For both the auto-tags and the user-tags, among each set of the top 20 TF-IDF weighted terms, some of them seemed to be more explicitly LULC-related, especially for the commercial recreation lands (CR) (refer to the bold terms in Tables 3 and 4). However, compared with those top 20 TF-IDF weighted auto-tag terms, it was more difficult to interpret the top 20 TF-IDF weighted user-tag terms. This was perhaps due to the heterogeneity of the user-tags. The contents of the user-tag terms could be nonphysical (e.g. geeky, wind, and early (Table 4)), while the auto-tags terms physically described the photo contents (e.g. ship, diver, and tiger (Table 3)). Therefore, in a way, the user-tag terms could be implicitly relevant to LULCs.

In addition, for both the auto-tags and user-tags, the 10 dimension loadings are highly different (Figure 5). The peak loadings of the 10 sets of weighted terms for the auto-tags across the 10 dimensions were concentrated in the first several dimensions, especially in the first dimension (Table 5 and Figure 6(a)), suggesting that the 10 types of the LULCs could not be effectively differentiated and delineated by the 10 sets of weighted terms. The peak loadings of the 10 sets of weighted terms for user-tags across the 10 dimensions seemed relatively more distinguishable (Table 6 and Figure 6(b)). However, certain sets of the weighted terms had multiple dimensional peaks with very approximate loading values. This was probably because, for both the auto-tags and user-tags, there were influential overlaps among the tags retrieved from the different types of LULCs, although the common tags were discounted by the TF-IDF method. This was further confirmed by the generally high pairwise semantic similarities, with an exception for the agricultural lands (A) being compared with the other types of LULCs (Figure 7). There might be
Table 3. The top 20 weighted auto-tag terms, for the 10 types of LULCs. The bold terms appear to be more explicitly relevant to the corresponding LULC.

| Rank | R     | I     | T     | C     | P     | A     | W     | O     |
|------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1    | Lens  | Wheel | Jetliner | Locomotive | Shelf stacks | Dolphin | Lunar | Duckling | Ship       |
| 2    | Wheel | Rampart | Ship | Riding | Riding | Wheel | Flamingo | Field mustard | Flamingo | Bowling |
| 3    | Camera | Rotunda | Jamb | Jettliner | Tulip | Bear | Team sport | Fron | Basket | Iceberg |
| 4    | Briefcase | Lunar | Airliner | Electronics | Office | Riding | River | Orchid | Diver | Geyser |
| 5    | hand Luggage | Door | Jet | Cap | Rampart | Tiger | Polar bear | Crag | Tulip | Rowing |
| 6    | Livestock | Softball | Wheel | Conference table | Riding | Cetecean | Tropical | Honeycomb | Bear | Brickwork |
| 7    | Sunflower | Airplane wing | Skyscraper | Roller skating | PANDA | Amusement park | Riding | Buffet | Backboard | Moth |
| 8    | Vault | Poodle | Apple | Salad | Rugby | Jockey | Shorebird | Ball | Heath | Ancient |
| 9    | Trellis | Skyscraper | Jacket | Performer | Spider | Team sport | Dashboar | Chocolate | Candy | Telephone pole |
| 10   | Poster | Plate | Spinnaker | Fieldwork | Bee | Panda | Ball | Cheerleader | Suit | Riding |
| 11   | Eagle | Bottle | Tropical | Freight car | Bread | Horse bridle | Snow | Berry | Night | Mist |
| 12   | Costume | Loft | Street sign | Sign | Wheel | Mammal | Tractor | Crane | Jet | Ruins |
| 13   | Barbecue | street Sign | Geometric | Wheel | Flamingo | Jaguar | Crab | Knitting | Lizard | Crew boat |
| 14   | Godfathers | High-rise | Surreal | Dashboard | Eye | Showjumpering | Mist | Cactus | Ball game | Branchlet |
| 15   | Ceramic | Rim | Team sport | Screen | Speckle dog | Team sport | Surreal | Elephant | Cave | Waterfowl |
| 16   | team sport | Geometric | Tortoise | Eskimo dog | Team sport | Food pan | Surreal | Diver | Giraffe | Bouquet |
| 17   | Trestle | Cockpit | Rowing | Food pan | Surreal | Crowd | Waterfall | Sailor | Kayak | Athlete |
| 18   | Girder | Roof | Amusement park | Wood joinery | Bird feeder | Giraffe | Surreal | Railroad track | Skyscraper | gorilla |

Table 4. The top 20 weighted user-tag terms, for the 10 types of LULCs. The bold terms appear to be more explicitly relevant to the corresponding LULC.

| Rank | R     | I     | T     | C     | P     | A     | W     | O     |
|------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1    | Dimples | Coolant | Picnic | Confirmation | Volunteers | Captivity | Brainstorm | Valleys | Canoe       |
| 2    | Samurai | Skyscrapers | Bisexual | Loked | Alley | Zoo | Fortune | Catering | Harbors | Grape |
| 3    | Beta | Swooping | Scanned | Presenter | Bop | Cetecean | Crack | Rattler | Thaw | Escape |
| 4    | Pentathlon | Sombrero | Drinking | Conveniences | Substantial | Whale | Environment | Savory | Con | Early |
| 5    | Roommate | Active | Meet | Husky | Volunteering | Captive | Bouldering | Folksinger | Dries | Cheerleader |
| 6    | Cutting | Customs | Wheels | Talk | Institution | Balsa | Magazine | Duo | Malarkey | Fonts |
| 7    | Breastfeeding | Skydiving | Stick | Reggae | Hospital | Conventions | Rhododendron | Hiking | Oiler | Sheep |
| 8    | Pregnancy | Oil | Reggae | Singlet | volunteers | Captivity | Brahms | Drums | Outrigger | Dinosaur |
| 9    | Swarm | Headlights | Best | Wresting | Show | Kill | Wind | Gowns | Published | Resorts |
| 10   | Kale | Outback | Athletics | Salsa | Association | Show | Amusement | Joystick | Diaries | Airbase |
| 11   | Modification | Volunteers | Network | Cars | Association | Show | Amusement | Joystick | Vampire | Beets |
| 12   | Racetrack | Leatherneck | Optics | Lumber | Presents | Dogs | Collar | Change | Graham | Coaches |
| 13   | Reunion | Butter | Cars | Airliner | Coastline | Tank | Photographer | Ants | Tomcat | Descent |
| 14   | Laws | Crocker | Bride | Heritage | Labor | Comic | Dragonfly | Brax | Gridiron | Sandals |
| 15   | Competition | Devils | Show | Male | Animal | Concert | Black | Brax | Tackle | Tackle |
| 16   | Blues | Dunes | Comics | Network | Festive | Trestles | Lib | Circuitry | Doy | Inclined |
| 17   | Students | Faces | Groom | Optics | Biological | Costuming | Preserve | Costume | Corpsman | Sculptures |
| 18   | Strand | Frosting | Cliffs | Corroded | Delivery | Theme | Ceremony | Davenport | Cruise | Bulb |
| 19   | Roadster | Sewn | Service | Windjammer | Laboratory | Imagery | Students | Delicate | Helicopters | Bushes |
several explanations for the overlaps. First, the positioning errors of photography devices could result in the overlaps. For example, photos that were taken in residential areas could be wrongly positioned in commercial areas. Second, the tags associated with a photo might not pertain to the location where the photo was taken. That is, a user could take photos at a long distance (e.g. taking photos about a water body from a rooftop in a residential area). Third, certain tags indeed can be associated with multiple types of the LULCs, such as sky and tree.

It was further observed that the pairwise semantic similarities of the 10 sets of weighted user-tag terms were generally lower than those of the 10 sets of weighted auto-tag terms (Figure 7). Perhaps, although a user could take a photo at a long distance, the user-tags could describe the location where the user took the photo, if the user preferred so. On the contrary, the auto-tags only physically described the contents of the photos. Additionally, the number of auto-tag terms was far less (predefined and restricted) than the number of user-tag terms (post-assigned without any restriction) (Section 2.3). Therefore, the level of semantic overlaps among the 10 sets of auto-tags was generally higher (Figure 7), due to the limited number of terms that were available to describe the diverse world. Indeed, some unique loading peaks of the 10 sets of weighted user-tag terms across the 10 dimensions could be observed (Table 5 and Figure 6(a)).
Figure 6. Loadings of the 10 sets of weighted terms across the 10 dimensions for the auto-tags (a) and the user-tags (b), respectively.

Figure 7. Semantic similarities among the 10 sets of weighted terms based on cosine similarity which are classified using natural breaks (Jenks). The upper right half is for the auto-tags and the lower left half is for the user-tags.
4.2. Future work

Based on the research findings above, how to use those top TF-IDF weighted terms to support the LULC separation (or classification) will be explored in the future work. Ways to minimize the semantic overlaps, making the tags usable in the LULC classification, will also be explored. For instance, the Flickr APIs provide the focus distance of a photo, which may be used as an offset to obtain the correct location of the photo contents, rather than the location where the photo was taken.

Acknowledgments

We thank Dr Chiao-Ling Kuo for her help on the Flickr data collection. We also thank the anonymous reviewers and the editors whose comments and edits have greatly improved this article.

Funding

This work is supported by the European Union LandSense project with the project title “A Citizen Observatory and Innovation Marketplace for Land Use and Land Cover Monitoring”, instrument Horizon 2020 and call identifier SC5-17-2015, demonstrating the concept of citizen observatories as an innovation action.

Notes on contributors

Yingwei Yan is a postdoc researcher in the GIScience Research Group, the Institute of Geography of Heidelberg University. His research interests are related to volunteered geographic information, citizen GIScience, big data analysis, digital humanity, natural disaster and crisis management, environmental modeling, land use/land cover changes, and artificial intelligence.

Michael Schultz is a postdoc researcher in the GIScience Research Group, the Institute of Geography of Heidelberg University. His research interests lie in land monitoring, accuracy assessment, OpenStreetMap and osmlanduse.org, climate change adaption, remote sensing time series, and machine learning.

Alexander Zipf is a professor in the GIScience Research Group, the Institute of Geography of Heidelberg University. His research areas include volunteered geographic information, crowdsourcing, citizens as sensors, analysis and processing of geographic data, 3D-GIS, 3D city and landscape models, Web-based GIS, GeoWeb 2.0, location-based services and ubiquitous GIS, interoperability and standards, spatial data infrastructures, and integrating new methods from Geoinformatics & GIScience in geography.

References

Antoniou, V., J. Morley, and M. Haklay. 2010. "Web 2.0 Geotagged Photos: Assessing the Spatial Dimension of the Phenomenon." *Geomatica* 64 (1): 99–110.

Arino, O., J. J. Ramos Perez, V. Kalogirou, S. Bontemps, P. Defourny, and E. Van Bogaert. 2012. *Global Land Cover Map for 2009 (Globcover 2009)*, Bremen: European Space Agency (ESA) & Université catholique de Louvain (UCL), Pangaean. doi:10.1594/PANGAEA.787668.

Bégin, D., R. Devillers, and S. Roche. 2013. "Assessing Volunteered Geographic Information (VGI) Quality Based on Contributors’ Mapping Behaviours." Paper Presented at *the 8th International Symposium on Spatial Data Quality*, Hong Kong, China, May 30–June 1. doi:10.5194/isprsarchives-XL-2-W1-149-2013.

Bossard, M., J. Feranec, and J. Otahel. 2000. *CORINE Land Cover Technical Guide: Addendum 2000*. Copenhagen: European Environment Agency.

Cihlar, J., and L. J. M. Jansen. 2001. "From Land Cover to Land Use: A Methodology for Efficient Land Use Mapping Over Large Areas." *The Professional Geographer* 53 (2): 275–289. doi:10.1080/00330124.2001.9628460.

Dorn, H., T. Törnros, and A. Zipf. 2015. “Quality Evaluation of VGI Using Authoritative Data—A Comparison with Land Use Data in Southern Germany.” *ISPRS International Journal of Geo-Information* 4 (3): 1657. doi:10.3390/ijgi4031657.

Estima, J., C. C. Fonte, and M. Painho. 2014. “Comparative Study of Land Use/Cover Classification Using Flickr Photos, Satellite Imagery and Corine Land Cover Database.” Paper Presented at *the AGILE2014 International Conference on Geographic Information Science*, Castellón, Spain, June 3–6.

Estima, J., and M. Painho. 2013a. “Exploratory Analysis of OpenStreetMap for Land Use/Cover Classification.” Paper Presented at *the 2nd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, Orlando, Florida, USA, November 5–8. doi:10.1145/2534732.2534734.

Estima, J., and M. Painho. 2013b. “Flickr Geotagged and Publicly Available Photos: Preliminary Study of Its Adequacy for Helping Quality Control of Corine Land Cover.” Paper Presented at *the 13th International Conference on Computational Science and its Applications*, Ho Chi Minh City, Vietnam, June 24–27. doi:10.1007/978-3-642-39649-6_15

Fonte, C., M. Minghini, V. Antoniou, L. See, J. Patriarca, M. A. Brovelli, and G. Milcinski. 2016. "An Automated Methodology for Converting OSM Data into a Land Use/Cover Map." Paper Presented at *the 6th International Conference on Cartography & GIS*, Albena, Bulgaria, June 13–17.

Fonte, C., M. Minghini, J. Patriarca, V. Antoniou, L. See, and A. Skopeliti. 2017. “Generating Up-To-Date and Detailed Land Use and Land Cover Maps Using OpenStreetMap and GlobeLand30.” *ISPRS International Journal of Geo-Information* 6 (4): 125. doi:10.3390/ijgi6040125.

Foody, G. M., L. See, S. Fritz, M. Van der Velde, C. Perger, C. Schill, and D. S. Boyd. 2013. “Assessing the Accuracy of Volunteered Geographic Information Arising from Multiple Contributors to an Internet Based Collaborative Project.” *Transactions in GIS* 17 (6): 847–860. doi:10.1111/tgis.12033.

Fritz, S., E. Bartholome, A. Belward, A. Hartley, H.-J. Stibig, H. Eva, P. Maysaux, et al. 2003. *Harmonisation, Mosaicing and Production of the Global Land Cover 2000 Database (Beta Version)*, Brussels: EC-JRC.
