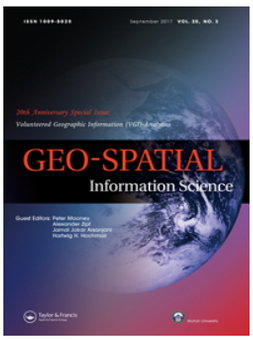

Geo-spatial Information Science

ISSN: 1009-5020 (Print) 1993-5153 (Online) Journal homepage: http://www.tandfonline.com/loi/tgsi20

Repurposing a deep learning network to filter and classify volunteered photographs for land cover and land use characterization

Lukasz Tracewski, Lucy Bastin & Cidalia C. Fonte

To cite this article: Lukasz Tracewski, Lucy Bastin & Cidalia C. Fonte (2017) Repurposing a deep learning network to filter and classify volunteered photographs for land cover and land use characterization, Geo-spatial Information Science, 20:3, 252-268, DOI: 10.1080/10095020.2017.1373955

To link to this article: http://dx.doi.org/10.1080/10095020.2017.1373955

GEO-SPATIAL INFORMATION SCIENCE, 2017
VOL. 20, NO. 3, 252–268
https://doi.org/10.1080/10095020.2017.1373955

Taylor & Francis
Taylor & Francis Group

Repurposing a deep learning network to filter and classify volunteered photographs for land cover and land use characterization

Lukasz Tracewski[a], Lucy Bastin[a] and Cidalia C. Fonte[b]

[a]School of Engineering and Applied Science, Aston University, Birmingham, UK; [b]Department of Mathematics, INESC Coimbra, University of Coimbra, Coimbra, Portugal

ABSTRACT

This paper extends recent research into the usefulness of volunteered photos for land cover extraction, and investigates whether this usefulness can be automatically assessed by an easily accessible, off-the-shelf neural network pre-trained on a variety of scene characteristics. Geo-tagged photographs are sometimes presented to volunteers as part of a game which requires them to extract relevant facts about land use. The challenge is to select the most relevant photographs in order to most efficiently extract the useful information while maintaining the engagement and interests of volunteers. By repurposing an existing network which had been trained on an extensive library of potentially relevant features, we can quickly carry out initial assessments of the general value of this approach, pick out especially salient features, and identify focus areas for future neural network training and development. We compare two approaches to extract land cover information from the network: a simple *post hoc* weighting approach accessible to non-technical audiences and a more complex decision tree approach that involves training on domain-specific features of interest. Both approaches had reasonable success in characterizing human influence within a scene when identifying the land use types (as classified by Urban Atlas) present within a buffer around the photograph's location. This work identifies important limitations and opportunities for using volunteered photographs as follows: (1) the false precision of a photograph's location is less useful for identifying on-the-spot land cover than the information it can give on neighbouring combinations of land cover; (2) ground-acquired photographs, interpreted by a neural network, can supplement plan view imagery by identifying features which will never be discernible from above; (3) when dealing with contexts where there are very few exemplars of particular classes, an independent a posteriori weighting of existing scene attributes and categories can buffer against over-specificity.

ARTICLE HISTORY
Received 8 April 2017
Accepted 19 June 2017

KEYWORDS
Land cover; land use; volunteered geographic information (VGI); photograph; convolutional neural network; machine learning

1. Introduction

In recent years, there has been an explosion in the popularity and prevalence of spatial data generation by citizens, through active collection initiatives such as OpenStreetMap, games and citizen science projects which tackle a wide range of topics, such as invasive species (Delaney et al. 2008), disaster response (Goodchild and Glennon 2010), cropland expansion (Fritz et al. 2012) and election violence (Meier 2008). This proliferation of data co-creation has been facilitated by the availability of cheaper sensors and GPS in smartphones. Web 2.0 technologies facilitate sharing, co-editing and online quality assessment of the generated information. Hand in hand with this active data generation is a rapid increase in the volume of voluntarily published resources, such as photos and reviews which are associated with some sort of locational information. Many terms have been coined to describe these types of data generated by the public but the term we will use to describe this particular mix of actively and passively published spatially referenced data is volunteered geographic information (VGI) (Goodchild 2007). One of the phenomena that may be mapped using such VGI sources (potentially in combination with more authoritative data) is land cover/land use. Geo-tagged photographs published to libraries, such as Flickr and Panoramio, are being increasingly investigated as potential sources of information in this context (Antoniou et al. 2016). If salient features can be identified and the position of the photographer is relatively certain, a subset of such photos may be useful for verifying and validating land cover/land use maps, and identifying changes in the landscape such as disturbance and vegetation change. In some cases, photographs may be presented to volunteers as part of a game (e.g. MissingMaps), which requires gamers to interpret relatively complex photographs to extract relevant facts about phenomena such as disturbance, agricultural practices or settlements (Fritz et al. 2012). Thus far, the

CONTACT Lucy Bastin l.bastin@aston.ac.uk

primary source of imagery for such applications has been aerial or satellite photography (e.g. detailed imagery from DigitalGlobe is used in the Tomnod platform) which offers a plan view of the ground. However, there is a potential role for volunteered photographs taken for entirely different purposes and published in repositories (e.g. Flickr) to fill the gaps, where no aerial imagery is available and to add significant value in terms of identifying key landscape features. The key challenge when exploiting such a vast and heterogeneous data source is to identify the most relevant and useful photographs from the deluge of available candidates, in order to most efficiently extract useful information while maintaining the engagement and interest of any volunteers assisting with classification. One option for automating this filtering process is to apply machine learning approaches such as deep learning to the content and metadata of the photographs, in combination with a user-defined set of priorities which define fitness for a particular purpose. The priorities of the original photographer submitting photos to Flickr will rarely align with those of a scientist trying to repurpose the image, so it is vital to identify the most salient images to avoid being swamped by irrelevant information.

This paper extends recent research into the usefulness of volunteered photos for land cover extraction, and investigates whether this usefulness can be automatically assessed in order to best focus citizen science efforts. We revisit a set of photos harvested from the web which were assessed for their usefulness by experts (Antoniou et al. 2016) and evaluates the degree to which the rule-based classification from the experts can be replicated by a neural network on the basis of the features identified in the photos. By repurposing an existing network which had been trained on an extensive library of potentially relevant features, we were able to carry out initial assessments of the general value of this approach, pick out features which were especially salient, and identify focus areas for future neural network training and development for this specific purpose. This approach also allowed us to test methods accessible to a general audience without specialized development and coding expertise.

2. Related work

2.1. Land cover and land use mapping—the context

Land cover or land use mapping is usually performed through the classification of satellite imagery. A variety of nomenclatures can be used in Land cover and land use mapping, some corresponding to land cover data and some also including land use information, such as CORINE Land Cover (European Environmental Agency 2006) or the Global Monitoring for Environment and Security Urban Atlas (UA) (European Environmental Agency 2012). The classification of land cover corresponds to the biophysical cover of the earth's surface, while land use is associated with arrangements, activities and inputs humans undertake for related to a certain type of land cover. The identification of different types of land cover using satellite imagery is easier than the identification of land use, since the latter often does not correspond to characteristics easily identifiable in aerial or satellite imagery using just reflectance. For example, a region covered with grass may correspond to several types of land use, such as sport fields, public or private parks or natural grassland. The opposite can also happen—a land use class, such as recreation areas, often includes several types of land cover.

Information about land use can be relatively easily provided by volunteers or be often easily identifiable in photographs taken at the earth's surface. Therefore, information provided by volunteers may be valuable, either when they are asked to identify land use classes directly, such as when creating vector data corresponding to land use information in OpenStreetMap, or by using photographs taken by the citizens. In the second context, the volunteers do not provide the land use/cover classes directly, but provide the data from which it can be extracted. The use of different nomenclatures raises problems when several sources of land use/land cover data are to be compared or combined. This requires the establishment of a mapping between nomenclatures. Even though this mapping is not always easy, several harmonization mappings are available for different nomenclatures (Arnold et al. 2013; Fonte et al. 2017).

2.1.1. Volunteered photos—the context

A thorough summary of online photo repositories and their protocols can be found in Antoniou et al. (2016). Some (currently relatively small) repositories focus specifically on land cover and land use; among these are the Degree Confluence Project and the Field Photo Library (Xiao et al. 2011). These data sources are well used for environmental modelling and validation (Foody and Boyd 2012; Iwao et al. 2006; Leinenkugel et al. 2014). These data can be assumed to be of interest for the purposes of this research. Therefore, we focus on expanding and supplementing this resource by repurposing and filtering public photographs from other domains, exploiting online repositories such as Flickr, Panoramio, Geograph and Instagram. The above-mentioned repositories host billions of photographs. These content and metadata is increasingly being analyzed to draw inferences about human social behaviour, tourism and the urban environment. The pool of photographs is rapidly expanding, with around 2 million public photographs a day being uploaded to Flickr (Franck 2016) and 58 million per day to Instagram (StatisticBrain 2016). Naturally, as alternative photo publishing platforms within a commercial market ecosystem, the repositories

differ in their focus and the dominant themes of the pictures published. A rough idea of these differences can be gained by investigating the optional text tags with which images are annotated. In terms of the images published which are shared publicly.

2.1.2. Flickr

Flickr leans towards the art and landscape side of photography, with numerous comments and discussions centred on the techniques used to capture or process images: trending tags often include topics such as 'depth of field' and 'exposure'. Because of this focus, many of the submitted photographs address landscapes: the all-time most popular image tags include 'sunset', 'water', 'sky', 'nature', and 'tree'. While these themes would appear very relevant to the recognition of landscape features and land cover, it is important to remember that these landscapes are frequently long shots that give little information about the location of the photographer (i.e. the actual geotag) and may be substantially processed.

2.1.3. Panoramio

Panoramio has a fundamentally spatial focus, since it specifically aims to showcase images attached to specific locations. The acquisition of Panoramio by Google in 2007 enhanced this function by embedding Panoramio pictures directly into Google Earth and Google Maps. Photo volumes in Panoramio are more modest; an estimated 93 million photos have been uploaded to the Panorank repository, but images are frequently deleted as users replace them with better versions, so the current number is probably lower. Daily estimates are not regularly monitored, but are estimated at between 20 and 40 thousand images per day; however, the frequency of geotagging is far higher than with Flickr photos. Tags, unsurprisingly, also frequently include concepts relevant to landscape features, such as 'mountain(s)', 'nature', 'forest', 'river', and 'urban'.

2.1.4. Geograph

The Geograph Project is an initiative designed to collect representative images across a number of sampling regions, namely, Britain and Ireland, Germany and the Channel Islands. The goal is for participants to collect at least one image for every square kilometer of these regions. Photographs tend to include architectural features or characteristic landscapes, and location and view direction are associated with the picture, since it is obligatory to report the position of the viewer and of the subject.

2.1.5. Instagram

As a social media platform, which is increasingly used for viral marketing by companies and celebrities, Instagram's trending tags are (at the time of writing at least) dominated by terms, such as 'cute', 'selfie', 'fashion' and 'best friends'. While various APIs are available for consuming georeferenced Instagram feeds (for example, Esri's GeoEvent connector) it was felt that for the purposes of this work, the streams of photos would require too much filtering, and so we confined the analysis to Flickr, Geograph and Panoramio. This also allowed us to evaluate our results against the work of Antoniou et al. (2016) by re-evaluating a set of data previously analyzed in their work.

2.2. Deep learning

Efficient filtering and classification of photographs to extract information on land use or land cover require a computer program to understand abstract concepts related to the interpretation of scene content. By using neural networks (Schmidhuber 2015) it is possible for software to learn these rules from a training set, without having to handcraft features (i.e. to characterize the elements of a scene in mathematical form) and provide them as inputs to a model. Neural networks are used in contexts where it is necessary to derive a relationship between variables, and where there are some observations of that relationship which can be used to train the network. A neural network is trained by propagating input data through layers of nodes to produce output values, which are then compared to the 'truth' to assess the goodness-of-fit. The desired output values may be continuous or discrete. Neural networks are most commonly used to assign categorical labels on the basis of continuous multivariate input data – for example, to recognize a text character from a variety of metrics derived from a pixelated image. Weightings in intermediate nodes are used to transform the input values to output, and these are iteratively adjusted to optimize the fit of the model. This process, known as 'back-propagation', allows the characteristics of a specific data set to be learnt. Unless the training data is extensive, with good representation of all the types of features to be distinguished, it runs a high risk of 'over-fitting', so that the model cannot reliably be applied to novel data. Over-represented classes in the training data can act as 'attractors' and significantly bias the accuracy of the final model. This is an important constraint in the context of VGI, which tends to be spatially patchy and biased towards particular themes.

The hidden layer is instantiated with values (most often random) which are later refined through back-propagation. The layer is called 'hidden', since it is difficult to provide an interpretation of the weights and what the network has learned.

Once there are three or more hidden layers in the neural network, we usually consider it deep—hence the term 'deep learning'. Fully connecting all neurons of one layer with all neurons of the next layer can lead to very complex optimization problems. For example, if an image with resolution 1000 by 1000 pixels is submitted to the network with one pixel value on each input, it

effectively means 10^6 values on input. Connecting the input layer with a hidden layer of the same size generates 10^{12} parameters to optimize. Adding more layers not only will lead gargantuan complexity, but will also cause severe overfitting: the neural network will do an excellent job in handling the cases it was trained on, but very poorly on new data sets (Schmidhuber 2015; Srivastava et al. 2014).

Convolutional Neural Networks (CNNs) address these problems and offer significant improvements over previous approaches (Krizhevsky, Sutskever, and Hinton 2012). Their architecture takes into account the spatial structure present in images, and introduces between the layers of the network an additional series of 'convolutional layers', each focusing on a particular region of the image. In order to further improve efficiency, parameters are shared across the network. Thus, the detection of a particular type of feature (to take a simple example, a vertical edge), once 'learnt' in one region, can be detected wherever it occurs in the image. As with the visual cortex of many animals, there is some overlap between the regions into which the image is divided. This region-based approach still makes local connectivity between neurons much easier to maintain, and allows the network to learn increasingly higher levels of abstractions. This makes the layers much easier to train, while having good grounds in computer vision and exploiting the phenomenon of spatial autocorrelation in imagery: distant regions within images are rarely semantically connected, but salient feature types, such as the aforementioned vertical edges, can occur anywhere in an unfamiliar image. A remarkable feature of this class of machine learning algorithms is the ability to generalize well and significantly outperform other approaches when it comes to dealing with abstract problems (Krizhevsky, Sutskever, and Hinton 2012; Schmidhuber 2015). Teaching the neural network to recognize a feature such as forest simply requires that the algorithm is shown sufficient number of photographs depicting forest, without having to explicitly define what 'forest' is. What constitutes 'enough' depends strongly on the complexity and variety of what the program tries to learn, but will require at least a few hundred labelled images per class. The requirement for a large number of training examples, as well as the computational power required for processing, can present a significant challenge, and for this reason, we took advantage of a ready-made model, which is further described below.

2.3. Identifying land cover from photos

In recent years a number of interesting initiatives have involved volunteers in identifying land cover and land use from images. In some cases, these are images taken from space or from the air, for example, the Geowiki cropland mapping initiative which asked volunteers to solve conflicts in widely used classified land cover maps (Fritz et al. 2015), the identification of invasive species in Hawaiian forests, the assessment of disturbance in and around protected areas (Bastin et al. 2013) or the recent 'gamification' of validation of the GlobeLand30 product (Brovelli et al. 2016). This 'view-from-above' has parallels with the classic remote sensing approach to landscape characterization, but instead of relying on spectral signatures or backscatter characteristics, land cover and land use types are identified by their characteristic shapes and patterns, easily picked out by the human eye.

Less frequently, photographs taken at ground level are used to record or verify land cover and land use maps, and in these cases many other factors come into play: for example, the focal length, orientation and viewpoint of a photograph, the accuracy of its locational information and its currency (many users of photo-sharing platforms upload scanned postcards or historic photos). Antoniou et al. (2016) analyzed the types of metadata that may be available associated to geo-tagged photographs, and which are available for Flickr, Panoramio and Geograph. Among these are orientation, date of upload and acquisition, focal length, tags, descriptions, titles and information about the photographer. The metadata required (mandatory and volunteered) varies according to the initiative and therefore, the metadata available for the photographs varies with their origin.

Many analyses which using volunteered photos use information other than the image itself: for example, parsing and using the associated tags to identify features of interest and delineating the areas (sometimes fuzzily defined) which users see as belonging to a particular named location (Gao et al. 2014; Li and Goodchild 2012) or see as attractive (Hu et al. 2015). On occasion, information about the user's identity is used to map trajectories (Jankowski et al. 2010) and or identify "localness" in shared photos and tweets (Johnson et al. 2016). Antoniou et al. (2016) analyzed the availability of tags, descriptions and titles in a set of 1000 photographs from each of Flickr, Panoramio and Geograph in the London area (corresponding to a total of 3000 photographs). The content of the harvested resources was not analyzed in that study, but only their availability and the number of available tags and words (for the descriptions and titles). The results show that for Geograph, only 34% of the photographs had tags, and this number increased to 70 and 79%, respectively, for Flickr and Panoramio, though the mean number of tags was smaller for Panoramio than for Flickr. This shows that the use of tags to identify the content of the photographs may leave out of the analysis a large number of photographs which could be useful to extract information. Therefore, methods that allow the analysis of the photographs themselves, instead of just the associated metadata, are useful.

Visual feature matching may be performed to identify landmarks (Kisilevich et al. 2010) or group photographs (Kennedy et al. 2007), but identification of land cover or

human disturbance from the photos themselves is, at the time of writing, less frequently researched. Deep learning approaches are ever more widely used to generate maps from images: for example, the Facebook initiative to map settlement configurations across 20 countries using 14.6 billion DigitalGlobe images at 50 cm resolution (350 TB data), combined with census data (Gros and Tiecke 2016) or the work by Castelluccio et al. (2015) to delineate land use types from the characteristic features which may be seen in detailed imagery. Albert, Kaur, and Gonzalez (2017) have also recently successfully classified typologies of city neighbourhoods using deep learning approaches combined with satellite imagery from Google Maps.

However, the above initiatives rely on the classic plan that delineates features from above, leaving scope for extra information to be gained from images acquired at the ground level. The closest work to what is addressed in this paper is that from Leung and Newsam (2015) who derived a classification of developed vs. undeveloped land for the UK, using photographs harvested from Flickr and Geograph, and (Zhu and Newsam 2015) – a campus mapping exercise which derived eight land use types from volunteered photographs in combination with a shapefile representing the zones on site.

In the study by Leung and Newsam (2015), the challenge of providing enough labelled images to train the machine learning algorithm was resolved by inferring the label through natural language processing from a description provided by the user. As the authors note, user-supplied text for an individual image is often not sufficient to assign it to a class, and so 1 × 1 km tiles were used to group photographs, in order to model topics efficiently. The authors use handcrafted features, namely colour histogram, edge histogram and gist descriptors, to train their model for scene recognition. Zhu and Newsam (2015), on the other hand, took the same approach to classification by using an off-the-shelf model, AlexNet (Krizhevsky, Sutskever, and Hinton 2012).

3. Specific approach for this study

In this work, we aimed to assess how far an off-the-shelf model which had been trained on a variety of potentially useful features could be adapted for our needs. The goal is to derive useful labels for a land cover or land use context without the need for an extensive gathering of 'ground truth', development of significant amounts of code, or heavy computational training of the network. An equally important goal is to assess and evaluate the limitations of this 'off-the-shelf' approach, and to try to characterize those contexts and photograph types where it is less reliable. In this way, we aim to derive some guidelines for best practice in the use of pre-trained models for specific use cases in the exploitation of volunteered photographs.

The CNN used in this study, Places205-AlexNet (Zhou et al. 2014), was trained by its authors on almost 2.5 million photographs, this allowed it to achieve 50% accuracy on identifying 205 "scene categories" (this term is explained in detail below). It should be noted that the choice was made purely on the wide availability of the pre-trained models for this neural network architecture and could be replaced with more accurate models. Training of the algorithm is an iterative process that usually a requires very large number of passes over the complete data set, making it a computationally expensive process which necessitates a huge set of labelled samples for training, analogous to the 'ground truth' of remote sensing classifications. In addition to this significant investment of resources, the creation of such a model requires expertise in computer vision and machine learning. For that reason, a number of studies, including this one, focus on retrofitting existing models, rather than building models from scratch.

Automated classification of photographs is highly relevant for those applications which involve volunteers in games or campaigns to identify interesting features from photos, since many photos are irrelevant, and simply presenting all available material runs the risk of boring volunteers and causing them to disengage. Ideally, we would like to filter photos in order to:

(1) Identify photos which are irrelevant or non-useful, and discard them;
(2) Identify images from which land cover/land use can be quickly and reliably identified (for example some types of built environment), harvest the labels and discard the photos from further analysis;
(3) Identify candidate land covers in the vicinity of remaining images for verification by volunteers
(4) Identify challenging and interesting photos that a user may enjoy deconstructing to extract more information than a machine can do.

The developed methodologies were applied to two study areas; one is the region used in Antoniou et al. (2016) that is situated in an urban area of London, UK and the other is located northwest of Paris, France, in a region which covers part of central Paris (as far South as Notre Dame) but also extends Northwards to a region with low-density urban areas and predominance of agriculture and forest. Since the London area was predominated by built environment and man-made features, the Paris region was selected in order to extend the classification challenge to a wider variety of land covers and land uses.

Within this study we aimed to assess how far we could achieve several distinct goals, as follows:

(1) Automating the identification of photographs which are useful for land cover classification.

(2) Extracting any information which can be immediately derived about land cover/land use.
(3) Relating the neural network outputs to an existing land cover classification, to assess how far accepted classes can be identified from image features.

We drew on past research by Antoniou et al. (2016) and specifically aimed to replicate their rule-based classification of photograph usefulness with an off-the-shelf combination of tools and a simple allocation of weights to tags which could be performed by a domain expert with no particular computational experience. The goal was to achieve a comparable stratification of images with much less investment of expert time, since the original classification of usefulness involved 7 experts each classifying 3 thousand photographs – a slow and tedious task.

3.1. Selection and setting up of algorithm and model

The Places205-AlexNet neural network (Zhou et al. 2014) has nine layers: an input layer, seven hidden layers and an output layer. The output layer consists of 205 scene categories, such as abbey, bedroom or mountain; a full list is provided in Table 1. The model outputs a value for each category, indicating the probability that a photo belongs to a certain class. This capacity is the result of training on MIT's Places database (http://places.csail.mit.edu/), a set of 2.5 million images, each labelled with a scene category. A novel image classified with this pre-trained network will usually belong to many categories, with varying scores. The last hidden layer of the neural network also provides valuable information about the photo content: a set of 4096 values that, in combination, form a 'signature' of the image. They represent high-level

Table 1. Scene categories from the Places205 project (Zhou et al. 2014).

abbey	construction_site	inn/outdoor	river
airport_terminal	corn_field	jail_cell	rock_arch
alley	corridor	kasbah	rope_bridge
amphitheater	cottage_garden	kindergarden_classroom	ruin
amusement_park	courthouse	kitchen	runway
aquarium	courtyard	kitchenette	sandbar
aqueduct	creek	laundromat	schoolhouse
arch	crevasse	lighthouse	sea_cliff
art_gallery	crosswalk	living_room	shed
art_studio	cathedral/outdoor	lobby	shoe_shop
assembly_line	church/outdoor	locker_room	shopfront
attic	dam	mansion	shower
auditorium	dining_room	marsh	ski_resort
apartment_building/outdoor	dock	martial_arts_gym	ski_slope
badlands	dorm_room	mausoleum	sky
ballroom	driveway	medina	skyscraper
bamboo_forest	desert/sand	motel	slum
banquet_hall	desert/vegetation	mountain	snowfield
bar	dinette/home	mountain_snowy	staircase
baseball_field	doorway/outdoor	music_studio	supermarket
basement	engine_room	market/outdoor	swamp
basilica	excavation	monastery/outdoor	stadium/baseball
bayou	fairway	museum/indoor	stadium/football
beauty_salon	fire_escape	nursery	stage/indoor
bedroom	fire_station	ocean	subway_station/platform
boardwalk	food_court	office	swimming_pool/outdoor
boat_deck	forest_path	office_building	television_studio
bookstore	forest_road	orchard	topiary_garden
botanical_garden	formal_garden	pagoda	tower
bowling_alley	fountain	palace	train_railway
boxing_ring	field/cultivated	pantry	tree_farm
bridge	field/wild	parking_lot	trench
building_facade	galley	parlor	temple/east_asia
bus_interior	game_room	pasture	temple/south_asia
butchers_shop	garbage_dump	patio	track/outdoor
butte	gas_station	pavilion	train_station/platform
bakery/shop	gift_shop	phone_booth	underwater/coral_reef
cafeteria	golf_course	picnic_area	valley
campsite	harbour	playground	vegetable_garden
candy_store	herb_garden	plaza	veranda
canyon	highway	pond	viaduct
castle	home_office	pulpit	volcano
cemetery	hospital	racecourse	waiting_room
chalet	hospital_room	raft	water_tower
classroom	hot_spring	railroad_track	watering_hole
closet	hotel_room	rainforest	wheat_field
clothing_store	hotel/outdoor	reception	wind_farm
coast	ice_cream_parlor	residential_neighborhood	windmill
cockpit	iceberg	restaurant	yard
coffee_shop	igloo	restaurant_kitchen	
conference_center	islet	restaurant_patio	
conference_room	ice_skating_rink/outdoor	rice_paddy	

Table 2. Scene attributes from the Places205 project (Zhou et al. 2014).

sailing/boating	spectating	tiles	glossy
driving	farming	concrete	matte
biking	constructing	metal	sterile
transporting	shopping	paper	moist
sunbathing	medical	wood	dry
touring	working	vinyl	dirty
hiking	using tools	plastic	rusty
climbing	digging	cloth	warm
camping	business	sand	cold
reading	praying	rocky	natural
studying	fencing	dirt soil	man-made
training	railing	marble	open area
research	wire	glass	semi-enclosed area
diving	railroad	waves	enclosed area
swimming	trees	ocean	far-away horizon
bathing	grass	running water	nohorizon
eating	vegetation	still water	rugged
cleaning	shrubbery	ice	vertical components
socializing	foliage	snow	horizontal components
congregating	leaves	clouds	symmetrical
waiting	flowers	smoke	cluttered
competing	asphalt	fire	scary
sports	pavement	natural light	soothing
exercise	shingles	sunny	stressful
playing	carpet	electric lighting	
gaming	brick	aged	

features of the image and it is from these values that scene categories are derived. A user can use this penultimate layer to build their own classifiers, and the authors of the Places205 network did just this, creating a set of 102 "scene attributes" which we have also used in our study. The scene attributes consist of classes like 'ice', 'working' or 'trees', with a full list to be found in Table 2. All of the mentioned values i.e. scene categories, scene attributes and the output of the last hidden layer, are used for classification in this work.

Some of the scene categories and attributes have clear relevance to land cover and land use: for example, 'forest', 'rock arch', and 'ocean'. Others relate to materials and characteristics from which at least some inference may be made about the surroundings: for example, 'shrubbery', 'enclosed area', and 'concrete'. Many of the labels are related to human activities which might be carried out anywhere, or interpretations of a scene which do not allow any inferences to be drawn: for example, 'reading', 'stressful' or 'business'.

In order to assess the value of our approach, we processed the output of the neural network in two different ways, in order to assess whether acceptable results could be achieved with a lightweight strategy available to relatively non-technical users. The simple approach (referred to as user-weighting [UW]) involved the allocation of binary weights (0 or 1), to each scene category and attribute, representing their relevance to particular categories of user interest (e.g. the land cover 'agriculture'). Using this approach, all the scores for a photograph can be weighted and summed to achieve a score for each of the user-defined labels. The rationale for this approach is that:

(1) It is simple to apply, requiring only some investment of time by one or more experts and some simple post-processing in a spreadsheet;
(2) In theory, it should be proof against overfitting, since the weights are assigned independently of any image training set;
(3) It should allow land covers which are rare in the training set to be adequately identified, since a user has independently flagged the tags which they consider to be indicative of those land covers.
(4) It allows derivation of a score for all images in the set, unlike a training/validation approach which requires a portion of the data to be set aside.

The more technical approach (referred to as decision tree [DT]) was based on supervised machine learning and involved building and training decision trees on the results from the original classification—an approach which requires little computational resources in comparison to original network training, but one that still calls for specific expertise. Whereas in the UW approach we orchestrate rules ourselves, in the DT we allow the computer to find an optimal set for us. For the specific algorithm, we selected XGBoost (Chen and Guestrin 2016), an open-source software library that provides a distributed tree boosting framework. By using an ensemble of weak prediction models, it can capture general rules governing the system, without having to explicitly define the relationships between or the importance of parameters. The validity of the model was always tested with fivefold stratified cross-validation, while confusion matrices were generated with one of the folds. To avoid strong bias towards the training set, we used penalized classification and restricted the depth of the constructed decision trees (defined as the length of the longest path from a root to a leaf) to five.

For each of our goals, we build three models, one for each of the neural network outputs (scene categories, scene attributes and numbers from the penultimate layer). We start by running a pre-trained model (in our case AlexNet) on images we would like to classify and storing its output. The latter becomes input to DT and UW methods. DT, as any supervised method, also requires that we provide labels to the training data set. The advantage over neural network is that far fewer labelled examples are needed for the algorithm to achieve good results and also only a few hyperparameters are left to tune. Our source code repository (https://github.com/RSPB/CitizenSensor) contains code for classification of images using pre-trained AlexNet model (in particular, file *image_classifier.py*) and an example of how to run the DT method (*classification_example_with_xgb.py*).

The contrasting approaches were applied to a series of goals (which define our output), as follows:

(1) Goal 1: identifying human impact in a landscape

For this exercise, we defined five classes as follows:

- bi = Built environment, indoors.
- *b* = Built environment, may be indoors or outdoors.
- hf = Human feature (e.g. a bridge, railway line, fountain, windmill) May be placed in a natural landscape.
- hl = Human land use (e.g. agriculture, gardens, golf course). Landscape may be vegetated but human influence is expected to affect a substantial area of the scene.
- *n* = Natural environment (and note that *u* = unknown)

The reasoning is that these categories could be useful for studies of fragmentation, habitat disturbance, etc. Each scene category and attribute was given a weight of 1 if expected to be indicative of these classes. After summing all weighted scores, the 'winning classes' were determined using the following algorithm:

(1) Find classes with score above predefined threshold. The threshold is calculated as *n*th percentile of the highest scoring class. For our experiments, we used the 70th, 80th and 95th percentiles. The particular values were selected arbitrarily as means to tune the method.
(2) If 'bi' class was found, '*b*' was added, as 'built environment, indoors' is a subset of 'built environment'.
(3) If '*b*' class was found, 'hf' was added, since buildings and other features characteristic of the urban environment are man-made features.

One of the authors of this article labelled 965 photographs with classes, while on 242 occasions assigning also an alternative class when it was not clear which class should be assigned to the photograph. In the UW method, we marked prediction as successful if any of the classes predicted by the algorithm was present as either the class or alternative class selected by the human expert. The DT approach was stricter: we considered a result as a match only if it was exact, i.e. the class found by the model was the same as the class labelled by the expert (in other words, the alternative class was not taken into account here).

(2) Goal 2: filtering photos by usefulness, based on perceived land cover

Antoniou et al. (2016) assessed a group of photographs from the London area and asked a group of experts to look for nine different land covers within the photographs, as used within the Geo-Wiki project: tree cover, shrub cover, grassland/herbaceous, cropland, wetland, artificial surfaces, bare rock/barren surface, snow/ice and water. Based on the presence or absence of those land covers, the usefulness of the photographs for identifying the classes was determined by experts, applying a set of rules that determined what the answer should be in case of doubts (Table 3).

In the UW approach, weights of 1 or 0 were assigned to associate each of the 205 scene categories and 102 scenes attributes with zero or more of the nine land covers defined in the original study. For example, 'rainforest' is associated with the 'forest' land cover class, and 'living room'indicates that a photo was taken within the built environment, but 'praying' cannot reliably be associated with any particular land cover or land use. This task was performed by one of the authors, and the resulting weights were multiplied by the image scores and aggregated to give a score for each of the aforementioned nine land covers, yielding 0 or more land cover labels for each photo. On the basis of these labels, thresholds and decision rules which mimicked Table 3 were applied to label each photo as 'useful', 'maybe useful'; or 'not useful'. Results for the London photos were validated against the original expert consensus (Antoniou et al. 2016) and for the Paris set, a subset of 965 photos were labelled by one of the authors as a validation set.

For the DT approach, the land cover identification step was skipped and the same validation labels were used for training in a fivefold cross-validation approach, to assess whether a model could learn the criteria for usefulness directly.

Table 3. Rules used to assist in the classification of the photographs as useful, from Antoniou et al. (2016).

No.	Description
1	Land cover is only considered when it is within about 10 m of the photographer, to take into account positioning errors of the photograph. Thus, land cover types in the far distance should not be considered
2	If it is possible to see or infer with reasonable certainty what is at the photographer's footprint (even when the footprint is not visible), and there is only one possible class from the nine classes considered, choose "yes"
3	If more than one of the classes above can be assigned to the photographer's footprint vicinity (using the 10 m limit defined in rule 1), choose "maybe"
4	If there is no information about what may be at the photographer's footprint, e.g. an aerial or panoramic view, then choose 'no'
5	Individual trees are discounted regarding the dominant land cover (e.g. a tree in a grass field) unless one can infer from the photograph that there are many trees around
6	For vintage photographs, the answer is 'no', since the land cover may have changed (or the photograph may be incorrectly geo-tagged)
7	For snow that completely covers the surface (so it is unclear what the underlying land cover is), because the study area is in London, the answer should be 'no'. Here context is used, not only the photograph, because in the city of London it is known that no permanent snow cover exists
8	For photographs taken underground, i.e. in a metro station, the answer is 'no'. If the station is clearly above ground and there is no other land cover type within 10 m, then the answer is 'yes' (artificial surfaces)
9	Water frequently causes difficulties because in many cases it is not possible to unequivocally determine if the photograph was taken from a boat (then the answer should be 'yes'), on a bridge, or at the water vicinity. Then, if the water is identified to be within 10 m of the photographer, the answer is 'maybe'

Table 4. UA classes of levels 2 and 4 (European Union, 2011, https://cws-download.eea.europa.eu/local/ua2006/Urban_Atlas_2006_mapping_guide_v2_final.pdf).

Level 2		Level 4	
Code	Class name	Code	Class name
11	Urban fabric	1110	Continuous urban fabric
		1120	Discontinuous urban fabric
		1121	Discontinuous dense urban fabric
		1122	Discontinuous medium density urban fabric
		1123	Discontinuous low-density urban fabric
		1130	Isolated structures
12	Industrial, commercial, public, military, private, and transport units	1210	Industrial, commercial, public, military and private units
		1222	Other roads and associated land
		1223	Railways and associated land
13	Mine, dump and construction sites	1340	Land without current use
14	Artificial non-agricultural vegetated areas	1410	Green urban areas
		1420	Sports and leisure facilities
20	Agricultural areas, semi-natural areas and wetlands	2000	Agricultural areas, semi-natural areas and wetlands
30	Forests	3000	Forests
50	Water	5000	Water

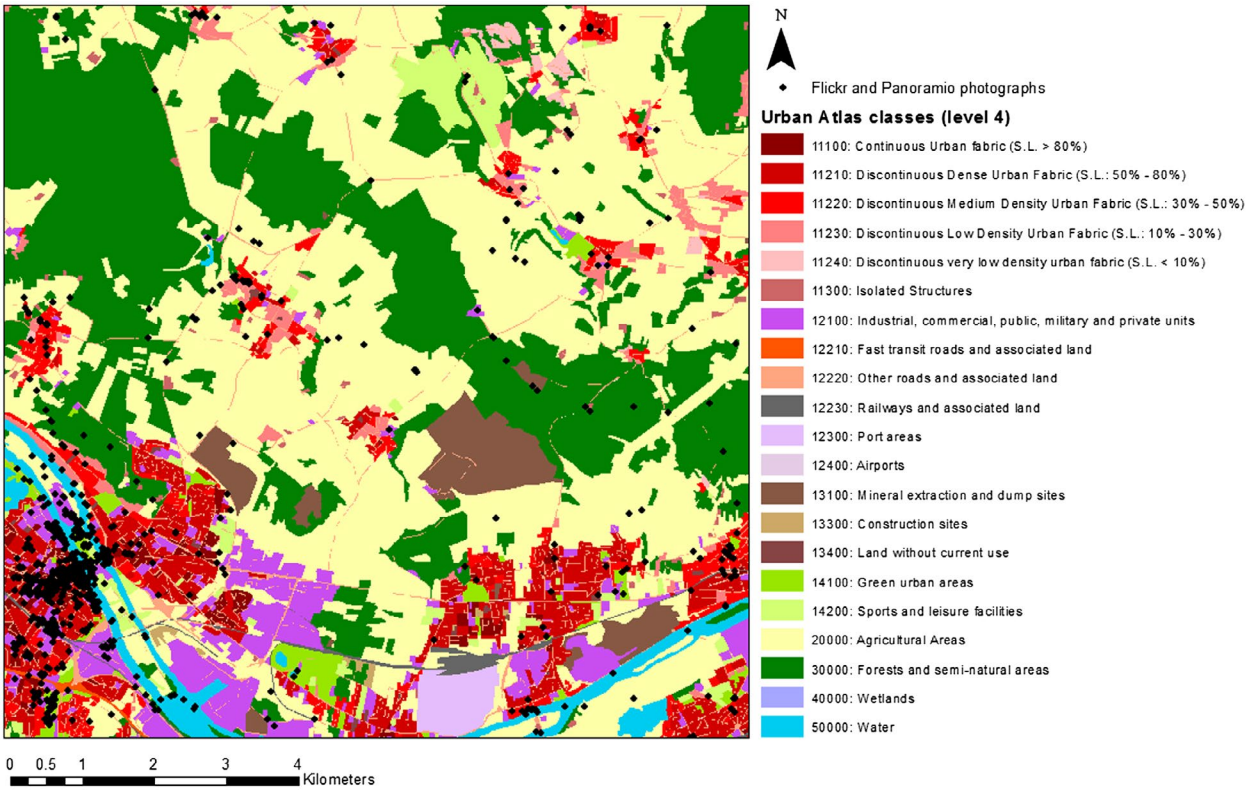

Figure 1. UA classes for the Paris study region, with locations of all photographs overlaid. The abbreviation "S.L" in the legend refers to the proportion of sealed surface which helps to define that class.

(3) Goal 3: identifying land cover as defined by UA

The European Environmental Agency's Urban Atlas (https://cws-download.eea.europa.eu/local/ua2006/Urban_Atlas_2006_mapping_guide_v2_final.pdf) is a high-resolution land use or land cover map of regions in Europe with more than 100 000 inhabitants. The created maps can be downloaded freely (https://www.eea.europa.eu/data-and-maps/data/urban-atlas). The created maps have a geometric scale of 1:10,000 and a minimum mapping unit of 0.25 ha for area features and 100 m for linear features. The nomenclature used in UA is organized into four levels of detail for the urban classes. Table 4 shows levels 2 and 4 of this nomenclature, as these were the ones used in this study. This goal aimed to determine if the classes associated with the photographs of the Paris region were correlated with the classes present in UA at the location of the photograph.

Figure 1 shows the locations of all the photographs harvested from the Paris study region, overlaid on the UA classification.

(4) Goal 4: identifying UA land cover classes in the surrounding area

In the process of discussing and evaluating the description of 'usefulness" in the above section, it became even more apparent to us that the goal of identifying land cover at the georeferenced point from

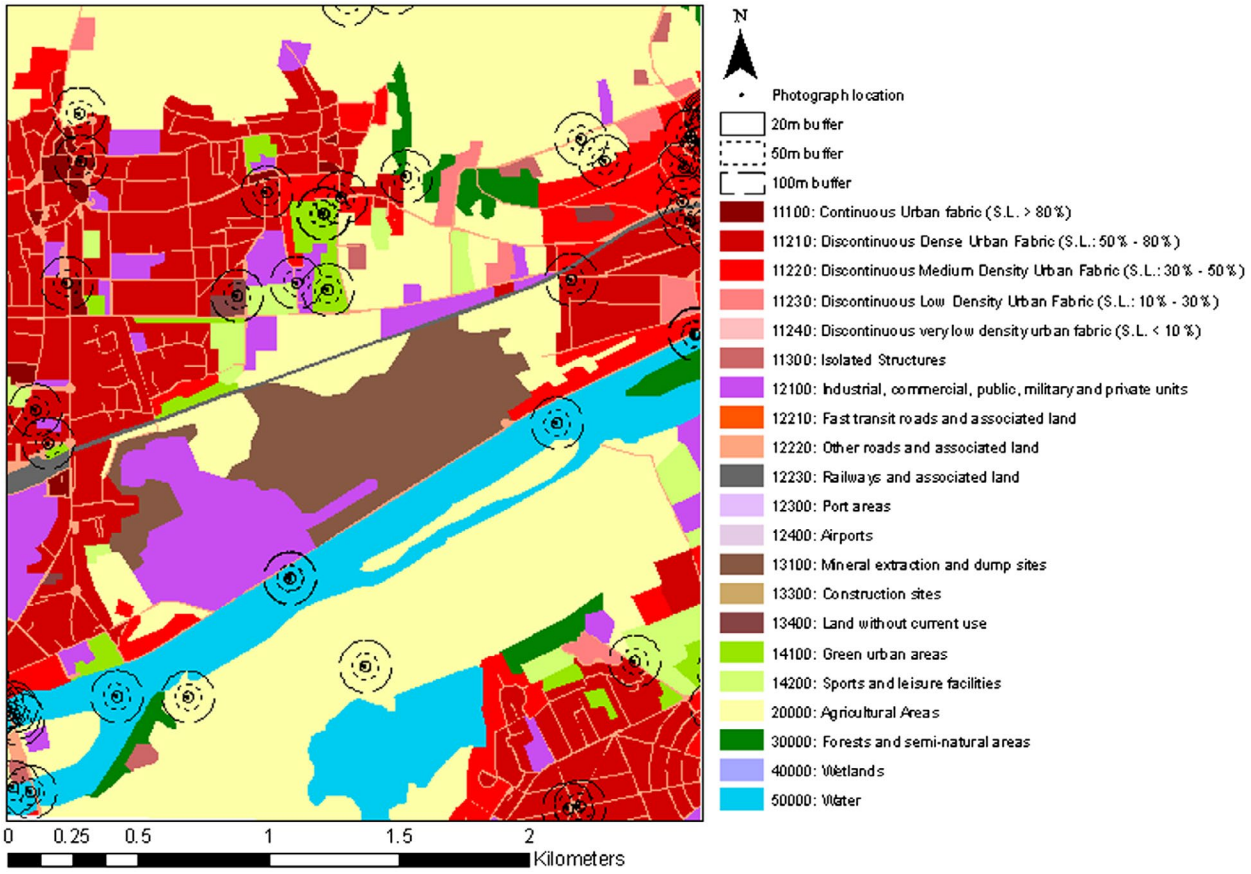

Figure 2. Example of 20, 50 and 100 m buffers around spots where photos were taken.

which a photograph is taken is highly challenging. For example, a photographer may be standing on a bridge while photographing water, or on a boat while taking a photograph of land. It is vanishingly rare that a photograph shared to a public image library will record a downward view of the location where the photographer is standing: rather, the images tend to represent transects or cones of vision of varying length which record a variety of the features in the neighbourhood, and which could form a useful complement to the 'plan-view' aerial photography and satellite imagery used for classical land cover classifications. Without information on depth of field and orientation (such as that explicitly recorded for the European Commission's LUCAS land cover survey and for the Degree Confluence project), it can be difficult to pin down the features located to exact locations. However, it is highly likely that future mobile phone development will make this easier, and that with sufficient volunteered photos, some useful triangulation could be done. In addition, the mix of features within a view may itself be informative in identifying land uses which have characteristic mixes of features: for example, built areas which have some level of vegetation. For this reason, we extended the above UA analysis to consider the land covers identified within 20, 50 and 100 m buffers around the reported location of the photograph. An example of these is shown in Figure 2.

Table 5. Frequency of UA level 2 classes in the data set at the specific location of each photo (point) and in the area around it (buffer).

Class	Point (%)	20 m buffer (%)	50 m buffer (%)	100 m buffer (%)
11	28.3	27.9	28.4	26.9
12	37.6	41.3	38.3	33.3
13	1.0	1.1	1.2	2.2
14	9.1	8.4	9.4	10.9
20	10.2	8.0	9.2	10.9
30	3.1	3.2	3.6	4.2
50	10.6	10.1	9.9	11.6

Table 6. Frequency of UA level 4 classes in the data set at the specific location of each photo (point) and in the area around it (buffer).

Class	Point (%)	20 m buffer (%)	50 m buffer (%)	100 m buffer (%)
1110	10.8	10.7	10.7	10.9
1121	10.6	10.9	13.6	14.1
1122	4.6	4.6	4.3	5.7
1123	1.7	2.0	2.2	2.0
1130	0.6	0.3	0.3	0.2
1210	20.2	15	14.2	14.7
1221	0	0.2	0.4	0.5
1222	16.4	28.6	26.7	21.8
1223	1.0	2.0	1.6	1.5
1340	1.0	0.7	0.6	1.2
1410	6.9	5.6	5.6	6.4
1420	2.3	1.9	2.6	3.0
2000	10.2	6.5	6.9	7.2
3000	3.1	2.6	2.7	2.8
5000	10.6	8.3	7.4	7.7

Some land covers, such as 'forest' had very little representation within the set of photographs, as shown in Tables 5 and 6. Others, such as 'urban fabric', were far more common. The extension of the area of interest around each photograph location by buffering altered the frequencies of the classes represented, but the relative rankings of the different classes remained roughly the same (Tables 5 and 6).

3.2. Result analysis

3.2.1. Goal 1: identifying human impact in a landscape

For the UW approach, performance improved as the threshold was lowered, allowing several alternative classes to be allocated to a photograph. Using a threshold of 0.7, an accuracy of 77.86% was achieved from the scene attributes, and an accuracy of 80.35% from the scene categories. However, this apparent success was largely an artefact of the lenient classification which, by permitting several class labels to be attached to each photo, increased the probability of a random match. A more stringent assessment of success was achieved when we used a threshold of 0.8, so that very few photos had an 'alternative class' and the measure of success was an exact match to the highest scoring class. Under these conditions, the accuracy dropped to 59.25 and 56.76%, respectively. Raising the threshold to 0.95 resulted in a further minor drop in accuracy (approximately 1%).

The DT approach, when trained and cross-validated on 480 labelled photos, performed much better (accuracy ranging from 74% for models using the penultimate layer of the neural network (scene attributes), to 65% with scene categories). An assessment of feature importance identified scene attributes such as 'open area', 'pavement', 'man-made', 'natural' and 'camping' as influential in the classification. However, there was one specific area where confusions were common using the DT approach: it disproportionately labelled photos of natural areas as having human features (Table 7). Average precision, recall and F1-score are presented in Table 8.

In the 'unweighted' variant, we calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account. By contrast, in the 'weighted' variant we calculate metrics for each label, and find their average, weighted by support (the number of true instances for each label).

The specific photos which were mislabelled in this way were identified and investigated, and are shown in Figure 3. It can be seen that, while some contain elements such as text or grey-scale colouring which could confuse a scene classifier, others clearly depict natural features such as woodland. On these photographs, the UW approach performed much better, classifying all as natural. A likely explanation is the relative paucity of 'natural' photos in the training set for the DT approach, meaning that the machine learning algorithm had relatively little information from which to learn the characteristics of this class. This class bias does not affect the UW approach. To some extent this bears out one hypothesized advantage of the UW approach: namely, that through exploiting independent expert opinion on associations between the off-the-shelf scene attributes and categories and the land covers of interest, it should be better at identifying land covers which are not sufficiently well-represented in the training set to be well learnt by the DT approach, and so should be particularly suited for scenarios where some classes are under-represented in the available training data.

Table 7. Confusion matrix, precision, recall and F1-score produced on a stratified test set, with number of photographs equal to 25% of all photographs in the Paris data set (the remainder were used for training the model) for the DT approach.

Actual class	Predicted class *b*	bi	hf	hl	*n*	*u*	Precision	Recall	F1-score
b	84	0	6	3	0	4	0.79	0.87	0.82
bi	4	10	0	0	0	0	0.71	0.71	0.71
hf	6	0	34	2	1	1	0.57	0.77	0.65
hl	9	0	7	31	0	2	0.86	0.63	0.73
n	0	0	9	0	4	0	0.67	0.31	0.42
u	4	4	4	0	1	12	0.63	0.48	0.55

Table 8. Average precision, recall, and F1-score for the DT approach.

	Unweighted	Weighted
Precision	0.70	0.73
Recall	0.63	0.72
F1-Score	0.65	0.72

3.2.2. Goal 2: filtering photos by usefulness based on perceived land cover

Even in the original study extended by this work, the conclusions were that usefulness defined by the given rules was very difficult to assess and agree upon. The goal of the original study was strictly to identify the land cover at the viewpoint of the photographer, which corresponds to the georeferenced location of the photograph. In the great majority of photographs this needs to be inferred with a reasonable certainty from the scene depicted, because the terrain underneath the photographer cannot be seen. This inference is subjective, and may produce variable results depending on the interpreter, especially since the land cover at the actual acquisition location may not correspond to the majority of the land use/cover information shown in the photograph. Therefore, potentially interesting and easily identifiable features within the field of view were often considered to be extraneous, or to lower the usefulness of the picture by adding uncertainty to the class that should be assigned to

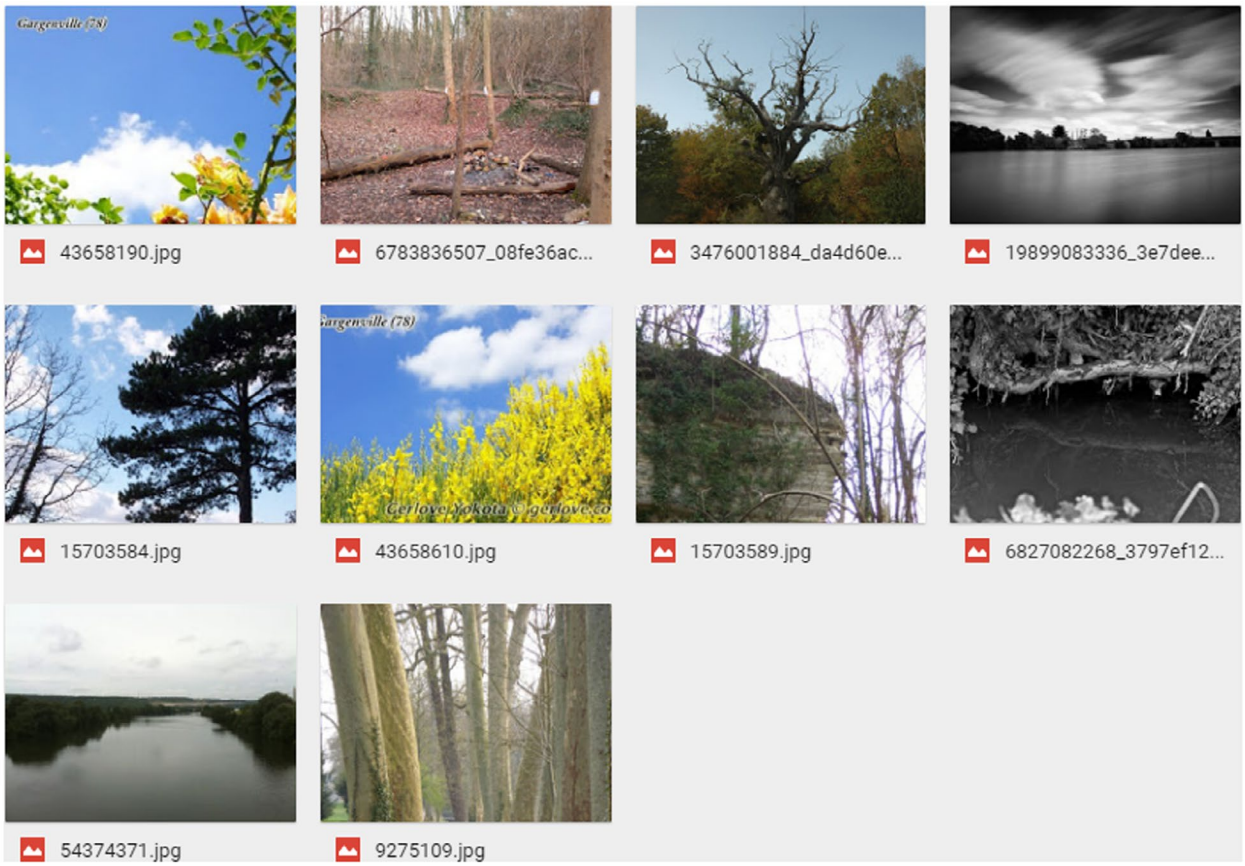

Figure 3. Photographs mislabelled by the DT method for identifying human impact in a landscape.

the exact location associated to the photograph. We had similar difficulty in deriving this particular definition of 'usefulness' by both classification strategies, since both UW and DT methods are specifically identifying and reporting features which may be at some distance from the photographer.

When trained and cross-validated with a fivefold approach on the London images, the DT approach achieved reasonable results (overall accuracy of 86%) but on the Paris data set it performed badly, with accuracies of 55%. This is unsurprising when we consider that the algorithm was trying to learn and generalize the rules behind several steps of human reasoning, some of them rather subjective. In trying to learn how the experts had classified 'usefulness', the DT approach assigned relative weighting to the scene attributes and categories. These generally correspond to highly informative factors in land cover identification: such as 'plaza', 'crosswalk', 'skyscraper', 'river', 'shopfront', 'formal garden' and 'field/wild' (all of which feature as key deciding factors in the derived classification).

By setting weights on the model, we could manipulate its sensitivity and specificity, and could control how strict the algorithm should be in rejecting photographs as not useful. In Table 9, we show an example in which we gave 'useful' images more weight to limit the incorrect rejection of pictures. This resulted in the acceptance for further analysis of 76 pictures which had been considered 'not useful' by the human expert, and 351 which had been labelled 'useful', meaning that in total, 427 images were proposed by the model as being useful for further interpretation, and only 53 were wrongly rejected.

Table 9. Confusion matrix, and the precision, recall, and F1-score for the "usefulness" prediction with the DT approach.

	Predicted				
Actual	No	Yes	Precision	Recall	F1-score
No	118	76	0.82	0.87	0.84
Yes	53	351			

Table 10. Confusion matrix, and the precision, recall, and F1-score for the "usefulness" prediction with the DT approach by tuning the weights such that we limit the number of false positives (increasing precision at the expense of recall).

	Predicted				
Actual	No	Yes	Precision	Recall	F1-score
No	168	26	0.91	0.68	0.78
Yes	128	276			

In this model, we put more weight on 'useful' images, thus limiting the number of false negatives, i.e. photographs that were falsely rejected by the algorithm as 'not useful'. By tuning the weights in the other direction (Table 10), we could reduce the number of images proposed by the model for further analysis to 302, but at the cost of rejecting a further 75 photographs which a human had labelled as potentially useful. In this model,

we put more weight on 'not useful' images, thus limiting the number of false positives, i.e. photographs that were incorrectly accepted by the algorithm as 'useful'.

The UW approach was specifically designed to build in the steps of land cover classification from the original paper and then to derive definitions of 'usefulness' from a strict application of the rules in Table 3. This approach performed acceptably on the London images (71.31% accuracy) but did not equal the performance of the DT algorithm. The UW approach performed particularly badly when extended to the Paris data set, with accuracies of 48% at most. This was little better than random, even when the analysis was reduced to two classes by collapsing the 'maybe' photographs into the 'not useful' class (Table 11).

3.2.3. Goal 3: identifying land cover as defined by UA

To address this goal, we again looked at scene categories, scene attributes and output of penultimate layer of the neural network. Using the DT approach, we built separate models for level 2 and level 4 UA categories for the Paris region, based on 1880 labelled photographs. In Table 12, we see that the results are quite poor, mostly due to having huge class imbalances and tiny numbers of samples on which to train many of the classes (Tables 5 and 6). Level 2 is better predicted than level 4, largely because of the aggregation of urban classes to higher level masks confusions between the varying mixes of built-up surface. Prediction accuracy of the selected predictors is shown in Table 13, with average precision, recall, and F1-score are presented in Table 14.

In general, the algorithm does a good job at recognizing built-up areas (for example, classes 11 to 14 from UA, Table 4) and distinguishing them from other classes. As can be seen from the related confusion matrix (Table 13), most often the algorithm confuses 'urban fabric' (class 11) with 'industrial, commercial, public, military, private and transport units' (class 12), which is hardly surprising given how semantically close they are. Also, higher error rates are present when it comes to misclassifying built-up areas as water bodies (class 50). In the majority of cases, this stems from the fact that in many urban photographs whose geolocation indicates a built land cover, a river is the object of interest (Figure 4). This tendency for photographs to include distant objects, which may, in fact, be informative, was the motivation for us to consider classes in the neighbouring vicinity in Goal 4.

Figure 4 was recognized by the DT method as class 50 (Water), though the UA label for the photograph's location is 12 (Industrial, commercial, public, military, private and transport units). Although the photographer was clearly standing on the bank of the river, the algorithm cannot infer this fact, and makes its classification based only on information present in the image itself.

3.2.4. Goal 4: identifying UA land cover classes in the surrounding area

One of the potential pitfalls of using labels which are co-located with a given photo is that they will not necessarily accurately represent the image's content. A photo of a forest can be taken from a wetland, which will result in misclassification: the algorithm will, correctly, recognize trees and predict class 'Forest' instead of 'Agricultural areas, semi-natural areas and wetlands'. To address it, we also considered the buffer around the spot from which the picture was shot. If we find that

Table 11. Confusion matrix, precision, recall and F1-score for the "usefulness" prediction with the UW approach.

	Predicted				
Actual	No	Yes	Precision	Recall	F1-score
No	411	164	0.56	0.53	0.54
Yes	180	206			

Table 13. Prediction accuracy of land cover with the DT approach on the extraction of UA classes from photographs, considering level 2 classes (7 classes) and level 4 classes (14 classes). "Chance of getting the result at random" is simply the chance of randomly guessing the class, 1/7th and 1/14th respectively.

Predictor	Level 2 Accuracy (%)	Level 4 Accuracy (%)
Scene categories	63.19	48.09
Scene attributes	62.77	50.43
Penultimate layer of the neural network	64.26	52.13
Chance of getting the result at random	14.29	7.14

Table 12. Confusion matrix for UA level 2 and the class prediction based on the scene attributes.

	Predicted class									
Actual class	11	12	13	14	20	30	50	Precision	Recall	F1-score
11	43	51	0	4	4	1	2	0.54	0.41	0.46
12	27	186	0	8	3	0	4	0.69	0.82	0.75
13	0	2	0	1	0	0	0	0	0	0
14	1	15	0	23	3	2	5	0.53	0.47	0.50
20	4	6	0	5	14	2	4	0.54	0.40	0.46
30	1	2	0	1	0	3	1	0.38	0.38	0.38
50	4	8	0	1	2	0	27	0.63	0.64	0.64

Notes: Here, 11 = Urban Fabric, 12 = Industrial, commercial, public, military, private and transport units, 13 = Mine, dump and construction sites, 14 = Artificial non-agricultural vegetated areas, 20 = Agricultural areas, semi-natural areas and wetlands, 30 = Forests, 50 = Water.

Table 14. Average unweighted and weighted average of the precision, recall, and F1-score for the level 2 UA classes presented in Table 11.

	Unweighted	Weighted
Precision	0.47	0.61
Recall	0.44	0.63
F1-Score	0.45	0.62

Figure 4. The example of urban photographs whose geolocation indicates objects which is not the objects of interest.

Table 15. Prediction accuracy of land cover with the DT approach for levels 2 and 4 of UA classes within 20, 50, and 100 m buffers defined around the photographs' locations.

	UA level 2			UA level 4		
	20 m	50 m	100 m	20 m	50 m	100 m
Predictor	buffer (%)	buffer (%)	buffer (%)	buffer (%)	buffer (%)	buffer (%)
Scene categories	86.38	91.28	94.04	61.49	68.09	80.00
Scene attributes	85.53	90.43	94.04	64.89	70.85	80.43
Penultimate layer of the NN	85.96	90.21	92.98	66.38	73.19	82.13
Chance of getting the result at random	25.88	30.69	38.59	16.36	21.62	29.29

the predicted class is among the classes within a buffer, then we consider it a match. Such approach significantly boosts accuracy, but also adds considerable odds that we will get a correct result purely by chance. However, the results achieved by the DT approach (Table 15) were still significantly above those which could be achieved by chance.

In Table 15, the higher chance of getting the result at random in a larger buffer is due to the larger number of classes that can be present in a given radius. The chance of getting the correct answer at random is calculated using the average number of classes per photo divided by the total number of classes.

3.3. Discussion

In the assessment of human impact, both DT and UW algorithms showed advantages and disadvantages, as they performed differently when applied to photographs from a region other than the one used for training. This shows that they may have different merits and may even complement one other.

When applied to assessing 'usefulness', the DT algorithm managed to capture essential rules behind the concept of 'usefulness' for the training set, but failed to generalize them to an independent set of images from another geographical location. We speculate that the method could still be considered valuable for certain scenarios in which experts classify a relatively small and evenly sampled set and then use it to train a model that is going to be used to filter for 'usefulness' a second set from the same area, possibly much larger. However, the rules defining 'usefulness' need to take into account the strengths and weaknesses of photographs acquired at ground level: for example, their potential for identifying features within a radius, or combinations of land covers which together constitute a specific land use. In addition, there are many cases where an automated algorithm cannot infer anything about the viewpoint of a photographer, while a human observer would be able to extract reasonably accurate information. A valuable cue in identifying these photos is the conflict between existing land cover labels and those assigned by the neural network. Such conflicts can be exploited by preferentially presenting these photographs to users for further interpretation.

There are many situations where a photograph can be labelled using an existing land use map, but the accuracy of the photo's location and the currency of the underlying map are in doubt. Ground-acquired photos whose acquisition date is known have a particular potential for identifying recent and dynamic changes which take time to filter into authoritative maps, and which may not be picked up by dynamic crowd sourced resources such as OpenStreetMap. In these cases, both the label assigned by automated overlay with existing maps and the features identified using a neural network could be combined as priors in more sophisticated Bayesian models which identify the key photographs for human interpretation, using contextual rules and apparent conflicts.

By tuning the DT algorithm, a user can define how strict in rejecting photographs the model should be. This is an important feature, since the risk or cost of false negatives and positives varies between use cases. For example, in a game-oriented photograph-identification application where plentiful pictures are available for a location, it is important to maintain user engagement by presenting them with only the most interesting and relevant pictures to interpret. By contrast, in a context where photographs are sparse but each one might contain critical information (for example, on earthquake damage or an invasive species) the cost of wrongly rejecting images as "non-informative" is much higher.

The application of the neural network to identify the land use / land cover classes used in UA on the photographs when only the photograph footprint was considered showed to have accuracies a little higher than

60% for level 2 classes with all approaches. However, if level 1 is considered (aggregating classes 11–14 into class 1–Artificial Surfaces) these results improve, as the built-up classes are fairly well distinguished from the other classes. As expected, the results are much better when buffers around the photographs are considered. These results were considerably better than what could have been achieved by chance, and as such they represent a significant first step towards identifying candidate land covers in an area.

4. Conclusions

Machine learning allows discovery of the rules which underlie a system. However, it comes at a cost: no matter how much one tries to avoid overfitting, the model will always represent the rules learned on the given training set. Some rules, like being able to determine a photographer's footprint even when it is not visible (Rule 2 in Table 3) are almost beyond reach for machine learning algorithms; the principles governing them are too abstract to learn, unless a significant number of training examples for the given case are presented to the algorithm during the training.

Currently, the pool of available photographs from Flickr, Panoramio and similar libraries is heavily biased towards tourism and towards heavily visited locations. Under these circumstances, photographs of some land covers for training a neural network are in short supply. The independent 'user weighting' approach tested in this paper shows some potential for buffering against this paucity of training material, since it exploits scene characteristics from a library which encapsulates extensive training (Zhou et al. 2014), and repurposes the available labels for particular contexts by allowing a user to weight their importance for their own particular use case. The results which could be achieved with very little technical ability, using an off-the-shelf tool, allow a rough filtering and classification of imagery which makes a significant contribution towards making sense of a vast and variegated resource.

That resource (i.e. the pool of publicly available photographs that can complement aerial and satellite imagery) will become much larger and more heterogeneous in the future. The recent success of the Pokémon Go game has demonstrated the potential to engage citizens in taking and sharing photographs from a variety of public spaces, and to direct players towards specific locations in order to gather 'evidence'. At the other end of the scale, inaccessible and relatively undisturbed areas are ever more frequently sampled by automated drones and camera traps, and there are moves towards more sharing and publishing of these image libraries (Constable et al. 2010; Cadman and González-Talaván 2014) so that they can be opened up for re-use beyond the identification of target animal species.

Both approaches had reasonable success in characterizing human influence within a scene, and in identifying the land use types (as classified by UA) present within a buffer around the photograph's location. There is some potential for refining this classification and using a transect-like approach for photographs where the field of view, orientation and direction are available, and this will be investigated in future work. In particular, we plan to apply these methods to land cover datasets containing systematically sampled photographs and transect information, in order to more rigorously assess the capabilities of a network specifically trained on such data.

As suggested in Antoniou et al. (2016), if the protocols to upload photographs or to comment on other users' photographs permitted volunteers to choose from a list of predefined tags associated with land use or land cover information, this metadata could be valuable to filter out photographs that had been designated as useful for this purpose. The list of tags to be defined should not be extensive, at least at an initial level, in order not to burden the volunteers with complex choices which might deter them from participating and documenting their contributions. The best list of tags to use, from the volunteers' and researchers' point of view, should be analyzed in future work.

We conclude that a neural network which is not specifically trained to identify land cover or land use can achieve modest levels of classification accuracy in isolation. Its outputs can be manipulated relatively easily to produce a useful 'first cut' at a classification and to pick out photographs which can be discarded either because the information they contain is easily extracted, or because they are likely to be irrelevant for the task at hand. In this way, we were able to use well-validated methods and benefit from a long and costly training exercise which, even though it was not designed for our specific field, yielded useful information on features that could be mapped to the composite land cover and land use types of interest for our context. The effort required was hugely reduced from the example of Antoniou et al. (2016), where considerable expert time and effort was invested to achieve a consistent labelling and classification. This initial study will help in focusing efforts in the planned future development and training of our own neural networks, which will be specifically tuned for existing labelled libraries of land cover and land use photographs. A particularly promising direction in which we propose to extend this work is the development of frameworks to combine varying types of evidence, exploiting the particular strengths of each. An excellent review of such fusion approaches for combining ground and overhead images in the land cover / land use classification context is provided by Lefèvre et al. (2017).

This will assist in focusing the efforts of human volunteers more valuably, and push forward the boundaries of citizen science by making the best possible use of the relative strengths of human and machines. This work represents a useful first step in evaluating the potential

of pre-trained neural networks for reuse in user-defined mapping contexts.

Acknowledgments

We would like to thank the support and contribution of COST Action TD1202 "Mapping and the Citizen Sensor" (http://www.citizen-sensor-cost.eu).

Funding

This work was supported by the COST Action [grant number TD1202] 'Mapping and the Citizen Sensor'.

Notes on contributors

Lukasz Tracewski is a PhD candidate in the School of Engineering and Applied Science at Aston University, UK. His main research interests are GIS and spatiotemporal analysis techniques.

Lucy Bastin holds a Senior Lectureship in the School of Engineering and Applied Science at Aston University, UK. She applies spatiotemporal analysis techniques to challenges in conservation planning, infection monitoring, and other environmental and socio-demographic contexts. Her work within the FP7-funded UncertWeb and GeoViQua projects addressed the management, reliable use and transfer of uncertainty information within a distributed, interoperable Model Web, and especially focused on standards which allowed users to augment and add value to the metadata for spatial resources. Dr. Bastin is currently on secondment to the Joint Research Centre of the European Commission where she is the lead developer on the Digital Observatory for Protected Areas (DOPA provides web-based validation and decision support tools for an international community of experts in biodiversity and forestry monitoring, and specifically supports the Convention on Biological Diversity).

Cidália C. Fonte is an Assistant Professor at the Department of Mathematics (Faculty of Sciences and Technology - University of Coimbra, Portugal). She is a researcher and member of the board of directors of the Institute for Systems Engineering and Computers at Coimbra. Her main research interests are a quality assessment of geographic information and uncertainty modelling, with applications in the areas of remote sensing, GIS, collection and use of volunteered geographic information. She was also a member of the Management Committee of the EU COST Action TD1202 "Mapping and the Citizen Sensor", where she chaired Working Group 4 dedicated to map validation. She holds a PhD in Geomatic Engineering.

ORCID

Lukasz Tracewski http://orcid.org/0000-0002-4778-4266
Lucy Bastin http://orcid.org/0000-0003-1321-0800
Cidalia C. Fonte http://orcid.org/0000-0001-9408-8100

References

Albert, A., J. Kaur, and M. C. Gonzalez. 2017. "Using Convolutional Networks and Satellite Imagery to Identify Patterns in Urban Environments at a Large Scale." Paper presented at The 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Nova Scotia, Canada, August 13–17. doi: 10.1145/3097983.3098070.

Antoniou, V., C. Fonte, L. See, J. Estima, J. Arsanjani, F. Lupia, and S. Fritz. 2016. "Investigating the Feasibility of Geo-tagged Photographs as Sources of Land Cover Input Data." *ISPRS International Journal of Geo-Information* 5 (5): 64. doi:10.3390/ijgi5050064.

Arnold, S., B. Kosztra, G. Banko, G. Smith, G. Hazeu, M. Bock, and N. Valcarcel Sanz. 2013. "The EAGLE Concept – A Vision of a Future European Land Monitoring Framework." Paper presented at The 33rd EARSeL Symposium, Towards Horizon 2020: Earth Observation and Social Perspectives, Matera, Italy, June 03–06.

Bastin, L., G. Buchanan, A. Beresford, J. F. Pekel, and G. Dubois. 2013. "Open-source Mapping and Services for Web-based Land-cover Validation." *Ecological Informatics* 14 (2): 9–16. doi:10.1016/j.ecoinf.2012.11.013.

Brovelli, M. A., I. Celino, M. E. Molinari, and V. Venkatachalam. 2016. "Land Cover Validation Game." Geomatics Workbooks Vol. 12 – FOSS4G Europe Como 2015. https://geomatica.como.polimi.it/workbooks/n12/FOSS4G-eu15_submission_197.pdf.

Cadman, M., and A. González-Talaván. 2014. "Publishing Camera Trap Data: A Best Practice Guide." https://gbif.org/resource/80927.

Castelluccio, M., G. Poggi, C. Sansone, and L. Verdoliva. 2015. "Land Use Classification in Remote Sensing Images by Convolutional Neural Networks." https://arxiv.org/pdf/:1508.00092v1.pdf.

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." Paper presented at The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17. doi: 10.1145/2939672.2939785.

Constable, H., R. Guralnick, J. Wieczorek, C. Spencer, A. T. Peterson, and The Vert Net Steering Committee. 2010. "VertNet: A New Model for Biodiversity Data Sharing." *PLoS Biology* 8 (2): e1000309. doi:10.1371/journal.pbio.1000309.

Delaney, D. G., C. D. Sperling, C. S. Adams, and B. Leung. 2008. "Marine Invasive Species: Validation of Citizen Science and Implications for National Monitoring Networks." *Biological Invasions* 10 (1): 117–128. doi:10.1007/s10530-007-9114-0.

European Environmental Agency. 2006. https://www.eea.europa.eu/publications/technical_report_2007_17.

European Environmental Agency. 2012. https://www.eea.europa.eu/data-and-maps/data/urban-atlas/mapping-guide/urban_atlas_2006_mapping_guide_v2_final.pdf/.

Fonte, C. C., J. A. Patriarca, M. Minghini, V. Antoniou, L. See, and M. A. Brovelli. 2017. *Volunteered Geographic Information and the Future of Geospatial Data: Using OpenStreetMap to Create Land Use and Land Cover Maps: Development of an Application*. Pennsylvania, PA: IGI Global.

Foody, G. M., and D. S. Boyd. 2012. "Using Volunteered Data in Land Cover Map Validation: Mapping Tropical Forests across West Africa." Paper presented at The IEEE International Geoscience and Remote Sensing Symposium 2012, Munich, Germany, July 22–27. doi: 10.1109/IGARSS.2012.6352675.

Franck, M. 2016. "How Many Photos Are Uploaded to Flickr Every Day, Month, Year." https://www.flickr.com/photos/franrckmichel/6855169886.

Fritz, S., I. McCallum, C. Schill, C. Perger, L. See, D. Schepaschenko, and M. Obersteiner. 2012. "Geo-wiki: An Online Platform for Improving Global Land Cover." *Environmental Modelling & Software* 31 (7): 110–123. doi:10.1016/j.envsoft.2011.11.015.

Fritz, S., L. See, I. McCallum, L. You, A. Bun, E. Moltchanova, and M. Obersteiner. 2015. "Mapping Global Cropland and Field Size." *Glob Chang Biol.* 21 (5): 1980–1992. doi:10.1111/gcb.12838.

Gao, S., L. Li, W. Li, K. Janowicz, and Y. Zhang. 2014. "Constructing Gazetteers from Volunteered Big Geo-data Based on Hadoop." *Computers, Environment and Urban Systems* 61: 172–186. doi:10.1016/j.compenvurbsys.2014.02.004.

Goodchild, M. F. 2007. "Citizens as Sensors: The World of Volunteered Geography." *GeoJournal* 69 (4): 211–221. doi:10.1007/s10708-007-9111-y.

Goodchild, M. F., and A. Glennon. 2010. "Crowdsourcing Geographic Information for Disaster Response: A Research Frontier." *International Journal of Digital Earth* 3 (3): 231–241. doi:10.1080/17538941003759255.

Gros, A. and T. Tiecke. 2016. "Connection the World with Better Maps." https://code.facebook.com/posts/1676452492623525/connecting-the-world-with-better-maps/.

Hu, Y., S. Gao, K. Janowicz, B. Yu, W. Li, and S. Prasad. 2015. "Extracting and Understanding Urban Areas of Interest Using Geotagged Photos." *Computers, Environment and Urban Systems* 54: 240–254. doi:10.1016/j.compenvurbsys.2015.09.001.

Iwao, K., K. Nishida, T. Kinoshita, and Y. Yamagata. 2006. "Validating Land Cover Maps with Degree Confluence Project Information." *Geophysical Research Letters* 33 (23): 265–288. doi:10.1029/2006GL027768.

Jankowski, P., N. Andrienko, G. Andrienko, and S. Kisilevich. 2010. "Discovering Landmark Preferences and Movement Patterns from Photo Postings." *Transactions in GIS* 14 (6): 833–852. doi:10.1111/j.1467-9671.2010.01235.x.

Johnson, I. L., S. Sengupta, J. Schöning, and B. Hecht. 2016. "The Geography and Importance of Localness in Geotagged Social Media." Paper presented at The CHI Conference on Human Factors in Computing Systems, Santa Clara, CA, USA, May 07–12. doi: 10.1145/2858036.2858122.

Kennedy, L., M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. 2007. "How Flickr Helps Us Make Sense of the World: Context and Content in Community-contributed Media Collections." Paper presented at The 15th ACM International Conference on Multimedia, Augsburg, Germany, September 24–29. doi: 10.1145/1291233.1291384.

Kisilevich, S., F. Mansmann, P. Bak, D. Keim, and A. Tchaikin. 2010. "Where Would You Go on Your Next Vacation? A Framework for Visual Exploration of Attractive Places." Paper presented at The 2nd International Conference on the Advanced Geographic Information Systems, Applications, and Services, Antilles, The Netherlands, February 10–16. doi: 10.1109/GEOProcessing.2010.11.

Krizhevsky, A., I. Sutskever, and G. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." Paper presented at The 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, December 03–06. doi: 10.1145/3065386.

Lefèvre, S., D. Tuia, J. D. Wegner, T. Produit, and A. S. Nassar. 2017. "Towards Seamless Multi-view Scene Analysis from Satellite to Street-level." *Proceedings of the IEEE* 99: 1–16. doi:10.1109/JPROC.2017.2684300.

Leinenkugel, P., M. L. Wolters, C. Kuenzer, N. Oppelt, and S. Dech. 2014. "Sensitivity Analysis for Predicting Continuous Fields of Tree-cover and Fractional Land-cover Distributions in Cloud-prone Areas." *International Journal of Remote Sensing* 35 (8): 2799–2821. doi:10.1080/01431161.2014.890302.

Leung, D., and S. Newsam. 2015. "Land Cover Classification Using Geo-referenced Photos." *Multimedia Tools and Applications* 74 (24): 11741–11761. doi:10.1007/s11042-014-2261-2.

Li, L., and M. F. Goodchild. 2012. "Constructing Places from Spatial Footprints." The 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, Redondo Beach, CA, November 06. doi: 10.1145/2442952.2442956.

Meier, P. 2008. "Crisis Mapping Kenya's Election Violence." https://irevolutions.org/2008/10/23/mapping-kenyas-election-violence/.

StatisticBrain. 2016. "Statisticbrain Instagram Company Statistics." https://www.statisticbrain.com/instagramcompany-statistics.

Schmidhuber, J. 2015. "Deep Learning in Neural Networks: An Overview." *Neural Networks* 61: 85–117. doi:10.1016/j.neunet.2014.09.003.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research* 15 (1): 1929–1958.

Xiao, X., P. Dorovskoy, C. Biradar, and E. Bridge. 2011. "A Library of Georeferenced Photos from the Field." *Eos, Transactions American Geophysical Union* 92 (49): 453–454. doi:10.1029/2011EO490002.

Zhou, B., A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. 2014. "Learning Deep Features for Scene Recognition Using Places Database." Paper presented at Advances in Neural Information Processing Systems 27, Montréal, Canada, December 08–13.

Zhu, Y., and S. Newsam. 2015. "Land Use Classification Using Convolutional Neural Networks Applied to Ground-level Images." Paper presented at The 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Washington, USA, November 03–06. doi: 10.1145/2820783.2820851.