Using social media, machine learning and natural language processing to map multiple recreational beneficiaries

Arjan S. Gosal[a], Ilse R. Geijzendorffer[c], Tomáš Václavík[b,c], Brigitte Poulin[d] and Guy Ziv[a]

[a] School of Geography, University of Leeds, Leeds, UK
[b] Palacký University Olomouc, Department of Ecology and Environmental Sciences, Olomouc, Czech Republic
[c] UFZ – Helmholtz Centre for Environmental Research, Department of Computational Landscape Ecology, Leipzig, Germany
[d] Tour du Valat, Research Institute for the Conservation of Mediterranean Wetlands, Arles, France

Declarations of interest: none.

Abstract

Information and numbers on the use and appreciation of nature are valuable information for protected areas managers. A promising direction is the utilisation of social media, such as the photo-sharing Flickr website. Here we demonstrate a novel approach, borrowing techniques from machine learning (image analysis), natural language processing (Latent Semantic Analysis (LSA)) and self-organising maps (SOM), to collect and interpret >20,000 photos from the Camargue region in Southern France. From the perspective of cultural Ecosystem Services (ES), we assessed the relationship between the use of the Camargue delta and the presence of natural elements by consulting local managers. Clustering algorithms applied to results of the LSA data revealed six distinct user groups, which included those interested in nature, ornithology, religious pilgrimage, general users and aviation enthusiasts. For each group, we produced high-resolution spatial and seasonal maps, which matched known recreational attractions and annual festivals in the Camargue. The accuracy of the group identification and spatial and temporal patterns of photo activity in the Camargue delta were evaluated by local managers of the Camargue regional park. This study demonstrates how Protected Area managers can harness social-media to monitor recreation and improve their management decision making.

Keywords: cultural ecosystem services, machine learning, self-organising maps, social media, recreation, beneficiaries.

Introduction

Cultural services, such as recreation, are the most challenging group of Ecosystem Services (ES) to study, as it is evident from their low frequency of inclusion both in scientific studies (Feld *et al.*, 2009) and national ecosystem assessments (Geijzendorffer *et al.*, 2017). Studies on identifying cultural services depend very much on the beneficiaries included (Martin-Lopez *et al.*, 2012; García-Nieto *et al.*, 2015). For example, García-Nieto et al. (2015) found that stakeholders with low and high environmental management influence had different perceptions of the spatial distribution of ES, including cultural services. As ES cannot exist in isolation from peoples' needs, understanding the linkages between ES and beneficiaries is vital (Haines-Young and Potschin, 2012; Nahlik *et al.*, 2012; Bagstad *et al.*, 2014). Due to the linkages between recreation and tourism, we refer to both as recreation in this study.

A variety of studies have investigated differences in recreation preference among different groups (Boxall and Adamowicz, 2002; Scarpa and Thiene, 2004; Arnberger and Eder, 2011; Gentin, 2011; Juutinen *et al.*, 2011; Ehrlich *et al.*, 2017). Results include different age quartiles showing a difference in the importance placed on several site attributes (i.e. the elderly placed more importance on activity type and litter, whereas the younger quartiles placed more importance on trail types and trail environment) on green spaces based in Vienna, Austria. A review from Gentin

(2011) suggests that ethnicity plays a significant role in recreation; with ethnic minorities preferring well-managed landscapes, with less preference for naturalistic environments.

The use of latent classes to identify differences in recreation preferences has shown that respondents characteristics can be used to group users into latent preference classes, for example on motivations for taking a trip and the stated preferences for wilderness park attributes (Boxall and Adamowicz, 2002). Latent class analysis (LCA) has been utilised by several studies in relation to recreational preference. Scarpa & Thiene (2004) found that climbers in North-eastern Alps could be placed in four classes, using variables including environment severity, the difficulty of climbs and shelter availability. Ehrlich et al. (2017) investigated recreational demand using perceptions towards water resource management in St. Johns River Basin (SJRB) in Florida (USA). They discovered two latent classes, both with similar demographic characteristics, though varying in attitudes and perceptions towards water management. In Oulanka National park (Finland), two latent classes of visitor type were identified, with nationality, income and time spent on site as significant variables for explaining membership (Juutinen *et al.*, 2011). Domestic low-income visitors who spent under 8 hours in the park characterised the first group, with the second being characterised by foreign high-income visitors who spent over 8 hours in the park (Juutinen *et al.*, 2011).

Preferential differences between stakeholders highlight the need for meaningful grouping of beneficiaries to understand and manage landscapes for their recreational needs efficiently. Whereas the above-mentioned studies demonstrate clear differences in “stated preferences”, few studies have looked at this topic from a “revealed preference” perspective, namely quantifying the spatial patterns of actual recreational activities of different groups, possibly because of the difficulty to conduct such studies with traditional survey methods. It has been previously shown that despite similarities between results in assessing cultural services at a landscape; where resources are limited, a revealed methodology is recommended (Hernández-Morcillo, Plieninger and Bieling, 2013; Milcu *et al.*, 2013; Gosal, Newton and Gillingham, 2018). Visitation data in Protected Areas (PAs) have been historically challenging to acquire, as their collection is time consuming, troubled by a variety of sampling issues and often competes with other research needs (Walden-Schreiner, Leung and Tateosian, 2018). However, these data are essential to develop strategies that minimise visitor impacts in PAs (Hadwen, Hill and Pickering, 2008; Walden-Schreiner, Leung and Tateosian, 2018). The monitoring of cultural services is particularly challenging to do at larger spatial scales because it excludes the use of specific common methods such as field survey. Despite environmental professionals seeing ES based approaches as being favourable (Martin-Ortega *et al.*, 2019), to inform managers of PAs, for example on the spatial pattern of different uses of the site, methods need to be feasible in terms of manpower and costs, coherent over time and cover a diversity of beneficiaries.

Billions of posts from millions of users are uploaded to social media platforms such as Facebook, Twitter and Instagram every year including geotagged images, videos or text (Hausmann *et al.*, 2017). Cost-effectiveness of using social media, or crowd-sourcing data, is a crucial driver for its uptake. Social media data is mostly free, in contrast to traditional methods of surveying which require greater human resources, and often incur trade-offs between detail and time available for the assessment (Richards and Friess, 2015; Hausmann *et al.*, 2017). Increased incorporation of Global Positional System (GPS), cameras and internet connection into smartphones and tablets have enabled many streams of scientific research (Di Minin, Tenkanen and Toivonen, 2015). Social media gives opportunities to access unstructured Big Data and is seen to be a “disruptive innovation”, allowing the progression of data-driven science (Kitchin, 2014). In recent years, there has been a concerted effort to utilise the power of social media to monitor tourism and recreational activities, highlighted by the growing body of studies using social media for assessing Cultural Ecosystem

Services (CES). The 'social-media-based method' is relatively new compared to other CES assessment methods such as direct observation and surveys (Cheng *et al.*, 2019). This has included preferences for biodiversity extracted from Instagram and Flickr, where Hausmann et al. (2017) found no significant difference compared to traditional surveys. The spatial distributions of images from Instagram for the City of Copenhagen have been found to show the main hotspots (Guerrero *et al.*, 2016). Image feature extraction on crowd-sourced data using a neural network has been used to ascertain outdoor elements that are found to be scenic (Seresinhe, Preis and Moat, 2017) and the use of geo-tagged photos from Flickr have been used in multiple studies as a proxy for visitation.

A recent study utilising geo-tagged images from Flickr (Sonter et al. 2016), investigated recreation in the conserved areas in Vermont, USA. They found eight predominant landscape attributes for visitation, including higher trail density, less forest cover and sites with more extensive areas. Tenerelli et al. (2016) investigated the role of variables that drive CES at a local scale in the Quatre Montagnes (bordering both the northern and the southern French Alps), capturing spatial fluctuations in preference. A study conducted across five sites across Europe by Oteros-Rozas et al. (2017) used photos uploaded to both Flickr and Panoramio to identify different cultural ecosystem service types. These included heritage, spiritual and social values being associated with wood pastures and grassland and anthropogenic landscapes, while recreation was found to be associated with mountain areas and water bodies (Oteros-Rozas *et al.*, 2017). A study in the Middle Atlantic Coastal Plain in North Carolina utilised the georeferenced social media images, content analysis and viewsheds to derive visual-sensory qualities of CES in the landscape (Van Berkel *et al.*, 2018). It was found that slope, water bodies and coastal attractions (including beaches) were important to the public, with agricultural areas being less valued.

Richards & Friess (2015) used photo content to classify areas by cultural use in urban mangrove sites in Singapore, finding recreation photos being more prominent around built sites and photographs of organisms in terrestrial and mangrove habitats. A later study in Singapore used automated image recognition approach to study the content and group photographs from social media and found the method was accurate and saved extensive periods of manual classification (Richards and Tunçer, 2018).

To date, none of these studies using geo-tagged social media datasets have considered explicit user groups in recreational preferences, despite research showing that social media data can be used to identify users to contribute to conservation science (Di Minin, Tenkanen and Toivonen, 2015). Despite the potential value of social data, managers of protected areas do not have the time or the capacity to analyse such data themselves and tools and algorithms will be needed to enable such applications.

Research into assessment of recreational ecosystem services should be aimed at improving comparability, whilst still maintaining context-specificity (Hermes *et al.*, 2018). Here, we demonstrate a novel method combining social media, machine learning and natural language processing, using a case study in the Camargue, France, which can be applied to other protected areas. The Camargue is used by many different actors, and the recognition of its cultural and natural heritage has resulted in its status as a Man and Biosphere reserve and a regional park. In the local management plan, it is stated that modelling the spatial and temporal dynamics can enable local stakeholders to understand the consequences of management decisions on the ecosystem functioning, biodiversity and services of the Camargue. Understanding the recreational usage of the area can allow for adequate planning, for example by setting up visitor infrastructures to increase public awareness or exclusion areas for the protection or development of sensitive ecosystems. The integration of the multifunctional uses of the park with its conservation objectives is very challenging and would benefit from information on the spatial distribution of the recreational use. Our specific objectives of this paper were to (a) identify a typology of users of the Camargue PA; (b) map the spatial and intra-annual pattern of use for each group of beneficiaries; (c) identify inter-annual and long-term trends in visitations for each group; (d) identify use value for protected area managers of the detected trends. We used a research and conservation organisation based within the study site for the validation of our results, which are produced by an automated algorithm and r analysis.

Methods

Study area

The Camargue Biosphere Reserve in the Rhône delta is a Ramsar site that covers natural habitats such as lagoons, brackish/freshwater marshes with emergent or aquatic vegetation, as well as halophilous scrubs and steppes. The studied area includes several protected area designations in the Biosphere Reserve (Figure 1), including the context of the surrounding land. These ecosystems, which are of significant importance for biodiversity (Heath *et al.*, 2000), are intermingled with agro-systems dominated by rice, an irrigated crop. The Camargue hosts a high species richness, typical of Mediterranean wetlands (Blondel et al. 2013). Wetland ecosystems of the Camargue are also essential for a range of ES such as climate regulation, flood mitigation, water purification, nutrient cycling, agriculture, fishing, cattle grazing, wildfowl hunting and bird watching. The functional biodiversity and habitats of the Camargue are predominantly influenced by the quantity and quality of water that is available year-round and large parts naturally dry up during the summer period.

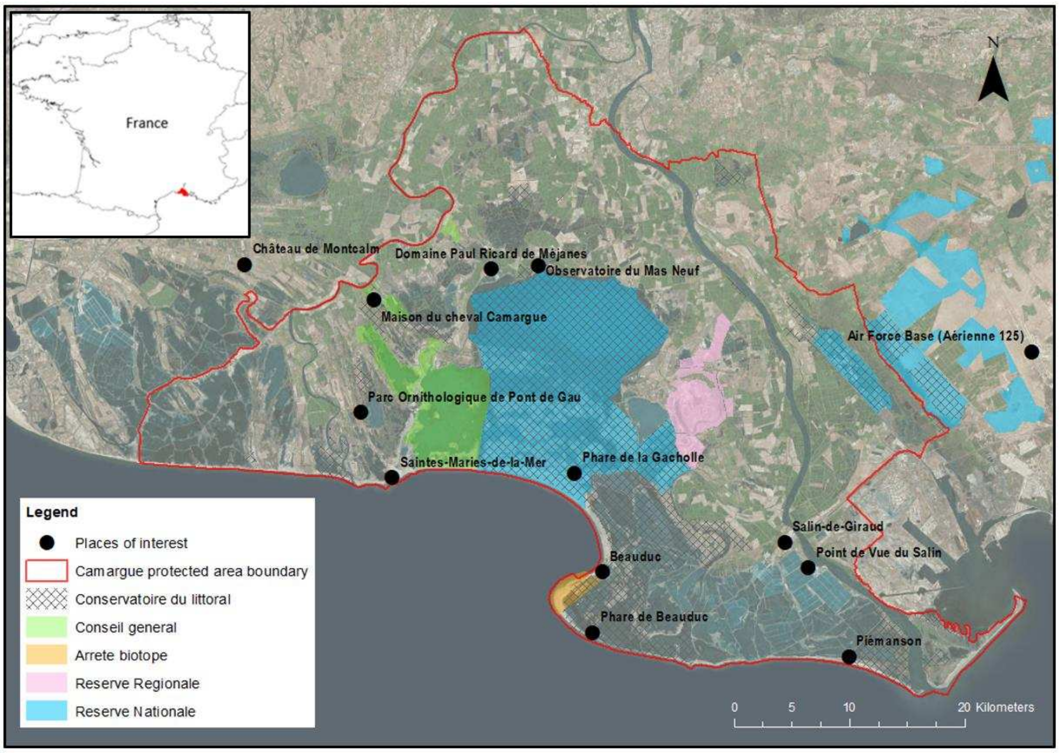

Figure 1: Satellite view of the study area of the Camargue with locations that could influence visitor photographs labelled. The Camargue protected area boundary is shown, with several protected areas highlighted. Map elements: © OpenStreetMap contributors, and the GIS community. Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS and Aerogrid.

The Camargue is covered by multiple labels and protection status which are partly overlapping. It includes a regional natural park, which is an inhabited rural area nationally renowned for its high, yet fragile, heritage and landscape values. A charter is defined collectively around the promotion and protection of traditional and natural heritage being endorsed by all the different sectors and actors in the Camargue delta. The responsible actors for the park management face the challenge of integrating multiple usages of ecosystem services and conservation of biodiversity in the same multifunctional area. Although land ownership is well known, it is much more difficult to obtain information on more spatial and temporal flexible use of the area, such as is the case for bird watching and tourism.

Photo Retrieval and Annotation

We retrieved images from the photo-sharing social media website Flickr using Python scripts and Flickr's Application Programming Interface (API). The images were downloaded with associated metadata, including longitude and latitude, date and time the photograph was taken, and the user ID of the photographer. A total of 20,051 images uploaded by 1292 users between 2007 and 2016 were downloaded. We used Google Cloud Vision, a machine learning algorithm for image analysis with the ability to detect labels, text, faces, landmarks, logos and image properties (Google Cloud Vision, 2017) that has been trained with extensive training sets.

The Google Cloud Vision API was used for automatic annotation of descriptive terms for the image's content with confidence scores, examples including 'flamingo', 'performing arts' and 'coast'. All images were analysed, with every annotation stored with its confidence score alongside the images' metadata. Images were not filtered, as this study does not purely look at the contribution of nature to recreation.

Typology of beneficiaries

Latent semantic analysis (LSA) was performed using the R language and coding environment (Wild, 2015; R Core Team, 2017). LSA is a technique used in this context to approximate meaning similarities between words or texts that are correlated with human cognitive phenomena such as semantic similarity (Landauer, Foltz and Laham, 1998). In LSA, latent semantic space is where 'documents' (in this study users are the 'documents') and 'terms' (herein image annotations) are represented as vectors, before applying local and global weighting and then calculating a singular value decomposition (SVD) is applied to a text matrix. Data were filtered to keep only those annotations with a Google Cloud Vision confidence score of ≥ 0.6. The LSA package was used to create a document-term matrix *M* (in this study this is a user by image annotations frequency table, an m x n matrix where *m (users)* = 1292 and *n (image annotations)* = 549) removing 'stop words' (commonly used words such as 'the'), with a minimum of 5 instances for each term to be included. The LSA was conducted using standard local weighting (log transform) and global weighting (inverse document frequency). The result was three matrices; T_k, S_k and D_k, where *T* is the term vector matrix, *S* is the singular values and *D* is the document vector matrix and where reduced dimensions $k = 82$. The dimensionality reduction in the LSA (the value for k) was calculated using the default 'fraction of the sum of the selected singular values to the sum of all singular values', with the default fraction of 0.5. This method uses a descending sequence of singular values for *s* and finds the first position where their sum is equal to or greater than the fraction specified. Hence 82 dimensions were outputted rather than 545.

The *S* matrix was plotted (see Figure S1), to discern the variance in the SVD and reduce the dimensionality further to reduce as much noise in the further analysis as possible, with an elbow in the plot found at $k = 6$. This method allows the identification of a point in the curve where the signal transitions to noise (Kutz *et al.*, 2016). The R *NbClust* package (Charrad *et al.*, 2014) was used with the 6 dimensions to calculate indicators for between 2 and 15 clusters to ascertain the optimal number of clusters to the reduced *D* matrix. Users were partitioned into this best number of clusters result (in this case six partitions) using the Ward-D algorithm (intra-cluster variation minimisation) (see Table S1). Word clouds of all terms for all images in a group were generated with R package *kohonen* (Wehrens and Buydens, 2007) to aid identification of the type of visitors group.

Seasonal Mapping of User Groups

Photo-User Days (PUD) is a measure that calculates the number of individual users that upload at least one photo on a unique day, in a particular location (Wood *et al.*, 2013), hence if a user uploaded five photos on one day, and ten on another day in the same location, the PUD would be 2. This avoids the problem of having users that upload many or few images from a single visit being counted differently. The number of PUD was calculated for each grid cell with a size of roughly 1 x 1 km across the study area. Seasonality was assessed distinguishing between the seasons (spring (March to May), summer (June to August), autumn (September to November), and winter (December to February)). The Flickr data was decomposed both by individual user groups and ungrouped data in R using the 'decompose' function using a multiplicative model. This decomposed data into the trend, seasonal and random components using moving averages (Supplementary Materials Figure S4). Maps for each group by season were created in ESRI ArcMap 10.3 to generate raster maps. Visitation area was calculated by summing the number of grid cells a user took photos in, on a single day, before being averaged across all visits per user and statistics calculated per group before calculating differences between groups using an ANOVA and post-hoc Tukey tests.

Spatio-temporal Patterns of Beneficiary Groups

We applied self-organising maps (SOM) to the mapped PUD data to identify the spatio-temporal patterns of use for all groups of beneficiaries. SOM is an unsupervised neural network, a competitive learning algorithm, uniquely suited for finding patterns in complex, high-dimensional datasets. It allows both (1) visualising complex data sets by reducing their dimensionality and (2) performing cluster analysis by grouping observations (grid cells in a map) into exclusive sets based on their similarity. Although caution is required when standardising input data and comparing outcomes of multiple model runs, the advantage of SOMs is that they depend less on expert rules or supervised threshold selection and are not restricted by the number of input features (variables and sample size). The quality of the data is vital for the quality of the outputs, with a larger PUD database giving more robust results.

As the SOM method is sensitive to outliers, we standardised the PUDs to zero mean and unit variance within each beneficiaries group. This Z-score standardisation also helps to interpret the results in terms of how much and in which direction the characteristic variable in each cluster deviates from the overall average. Optimum cluster size was determined using a Davies-Bouldin (DB) Index and mean distance to cluster centroids (see Supplementary information, Figure S5) calculated for a variety of cluster sizes ranging from 3 to 20 clusters. Identifying a natural break in both measures, we chose five clusters as they provided an optimal trade-off between the number of clusters and their quality of data representation. The SOM analysis was conducted using the *kohonen* R package (Wehrens and Buydens, 2007).

Validation by experts

An expert consultation was used to elicit local knowledge on user groups, with experts from Tour du Valat (Research Institute for the Conservation of Mediterranean Wetlands), i.e. co-authors B. Poulin and I. Geijzendorffer of this paper, located within the Camargue and with two representatives of the park management. During the consultation, the experts and local park management were presented with large format seasonal maps for each of the six identified groups, with associated word clouds. The experts evaluated and helped to validate the distinguished user groups and the patterns their photos described. Reactions and comments from the experts were considered to formulate and visualise the final results.

Results

Our analysis of Flickr photographs identified six distinct groups of cultural ES beneficiaries in the Camargue (Table 1). Together with local protected area managers, we interpreted these general visitor groups as two types of tourist groups, nature tourists and general tourists (1 and 2), (3) bird lovers, (4) equestrian enthusiasts, (5) aviation enthusiast and (6) religious visitors. Naming groups 1 and 2 were more challenging due to the similarities in the groups, though word clouds for group 1 featured more nature terms referring to fauna and flora, whereas group 2 included more frequently terms such as 'wall' and 'building'. The experts confirmed the patterns of group 1 and 2, but also felt that their patterns did not differ substantially. The average photos taken by users from each group varied from nature tourists averaging 1.60 to religious visitors averaging 50.95. Calculations of PUD showed aviation enthusiasts had the highest value of 7.50 with the lowest value for nature tourists. Interestingly, it was this latter group which made up the largest share of users (72.76%), and aviation enthusiasts having the smallest number of members (0.62%). This suggests that aviation enthusiasts are the most interesting in taking (and uploading) photographs despite being one of the smallest groups and covering one of the largest spatially distinct areas. The experts were able to confirm the existence of the visitors of the area taking pictures of planes, but they were surprised by the relative number of photos taken as well as the linear pattern they described in the Camargue landscape. The area visited by users in individual visits (based on their photos) was between 1.23 to 2.90 km^2, with equestrian enthusiasts covering the most extensive area. Post-hoc Tukey tests showed that the amount of area visited by each group was significantly different among nature tourists, bird lovers and equestrian enthusiasts and from the remaining groups (general tourists, aviation enthusiasts and religious visitors).

Group	Name and description	% of total users	Average photos per user	Average PUD per user	Area visited (average per user/per day in km²)			Tukey groups*
					Mean	Median	SD	
1	**Nature tourists** *Interests in the sea, shore, beach and nature.*	72.76	1.60	1.47	1.23	1.00	0.65	c
2	**General tourists** *Tourists who enjoy nature, though less interested in animals and more infrastructure such as human sites.*	8.90	6.50	4.23	1.98	1.64	1.61	bc
3	**Bird lovers** *Those with interest in taking photos of birds.*	12.62	6.74	3.77	2.11	1.33	1.87	b
4	**Equestrian enthusiasts** *Those with interests in horses and other mammals/wildlife.*	4.49	16.61	7.40	2.90	1.73	2.58	a
5	**Aviation enthusiasts** *Those with an interest in aircraft.*	0.62	10.95	7.50	2.00	1.00	2.45	bc
6	**Religious visitors** *Those who visit for pilgrimage, and the associated activities in spring in Saintes-Maries-de-la-Mer.*	0.62	50.95	4.63	1.34	1.00	0.69	bc

Table 1: Total numbers of users identified for each group using LSA, with average PUD per user (PUD/users), and area visited per user/per day. Total PUD was 2.33 per user (ungrouped). *One-way ANOVA determined statistically significant differences between the groups (F(5,1286) = 41.42, p < 0.0001). Tukey post hoc tests were used, groups not significantly different from each other are represented with the same letter.

PUDs were plotted by yearly season, and Seasonal Mann-Kendall tests were used to identify trends in the individual groups, with significant positive trends being seen in the visitation rates based on photos uploaded to Flickr for nature tourists, bird lovers, and equestrian enthusiasts. Ungrouped data also showed the same significant trend (Figure 2).

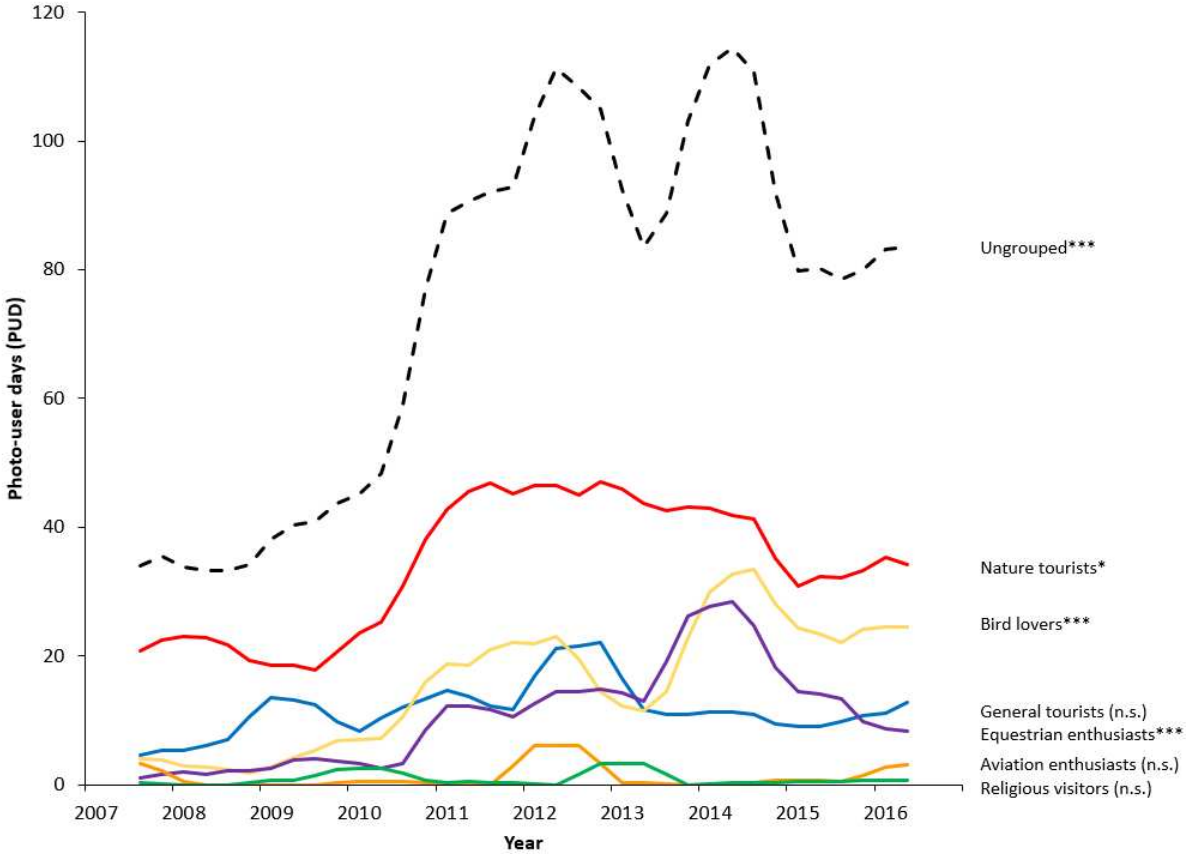

Figure 2. Trends in photo-user days (PUDs) across all groups decomposed by season (Supplementary Information, Figure S4), asterisks denote a significant trend using Seasonal Mann-Kendall tests (* $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$) or non-significant (n.s.) (Supplementary Information, Tables S2).

All PUDs were mapped for ungrouped data (Figure 3) and individual groups (Figure 4-9) to provide an indication of visitation and recreation. The ungrouped data showed a distribution of pixels across the landscape with aggregations around the coast near the western village of Saintes-Maries-de-la-Mer and past Phare de la Gacholle and Beauduc to Salin-de-Giraud, Port Saint-Louis du Rhône and Piemason in the east of the Camargue. Photos were distributed across the remaining landscape, including visual clustering around the edge of the Vaccarès lagoon. The highest visitation was at Parc Ornithologique de Pont de Gau and Saintes-Maries-de-la-Mer throughout spring, summer and autumn, with the latter being the sole hotspot in winter. The general spatial pattern of the photos was considered very logical by the experts, as much of the area in the Camargue is in private hands or are protected nature areas. The photos taken by tourists therefore clearly show the accessible areas (e.g. beaches, towns, visitor centers) as well the scenic look out points accessible from the road.

Nature tourists (Figure 4) had the most substantial amount of visitation pixels from all groups. PUDs are strongest in the spring and summer, with a hotspot again at the village Saintes-Maries-de-la-Mer. Visitation is expansive and follows the coast and the periphery around the lagoon. General tourists (Figure 5) have a reduced visitation pattern across the landscape compared to nature tourists though follows a similar pattern.

The most frequented area by bird lovers was around the Parc Ornithologique de Pont de Gau and the neighbouring lagoon. Scamandre Regional Nature Reserve, in the north-west of the Camargue, has also many trails attracting bird-watchers. For equestrian enthusiasts (Figure 7) in the spring, the spatial patterns follow the areas around Saintes-Maries-de-la-Mer, the lagoon, and the beach areas of Beauduc and Piemanson and Salin-de-Giraud. This continues in the summer, though a noticeable amount of visitation occurs to the north-west of Saintes-Maries-de-la-Mer around a double bend of a tributary of the River Rhone where a horse rental service (Ballade a Cheval) is located. Another noticeable aggregation is observed at the Maison du Cheval Camargue, or House of Horses, in winter. This protected estate of 287 ha located West of the Vaccarès lagoon, holds championships and other activities for Camargue horse enthusiasts.

Aviation enthusiasts (Figure 8) are mostly visiting an Air Force base located East of the Camargue in the summer months, with visitation following a north-west to south-east spatial pattern. Religious visitors (Figure 9) are spatially aggregated around the village of Saintes-Maries-de-la-Mer, with the frequency of terms (Supplementary Information, Figure S3) inferring the pilgrimage that attracts a lot of visitors every year around Easter.

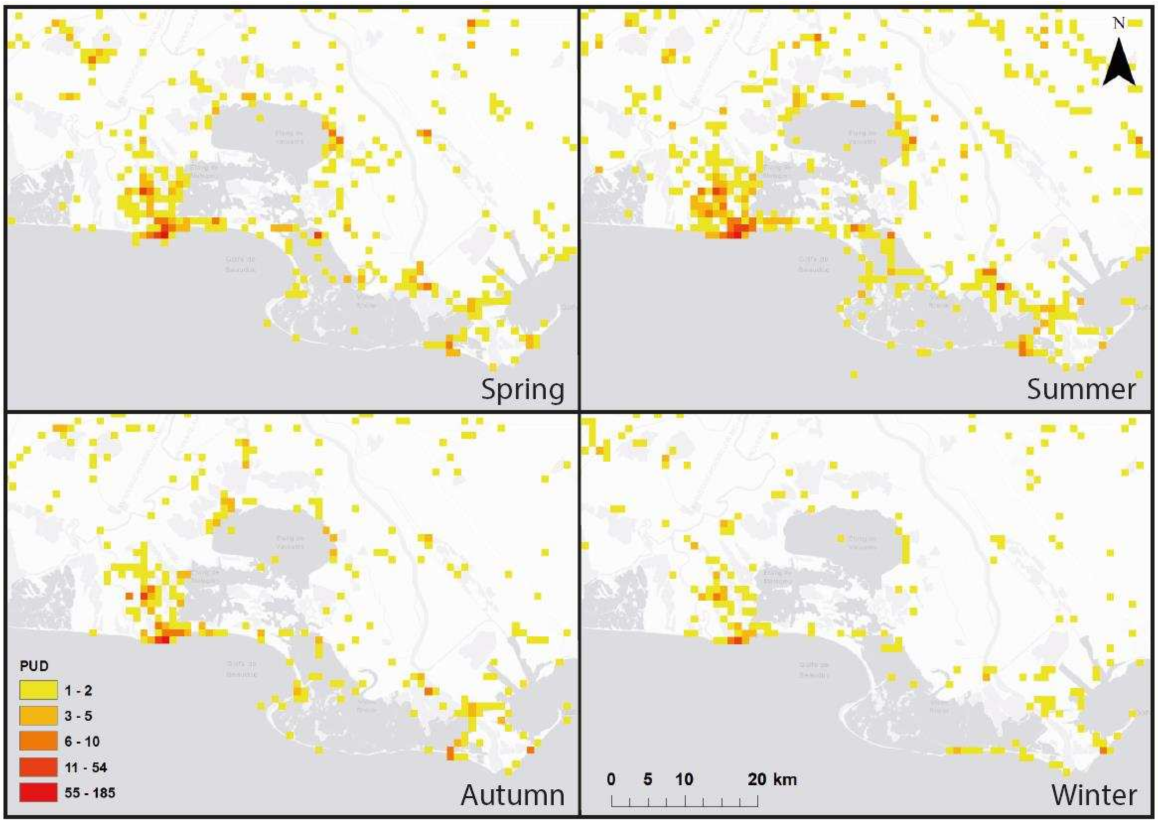

Figure 3: Seasonal distribution of Photo-User Days (PUD) in the Camargue. Mapping elements: Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community.

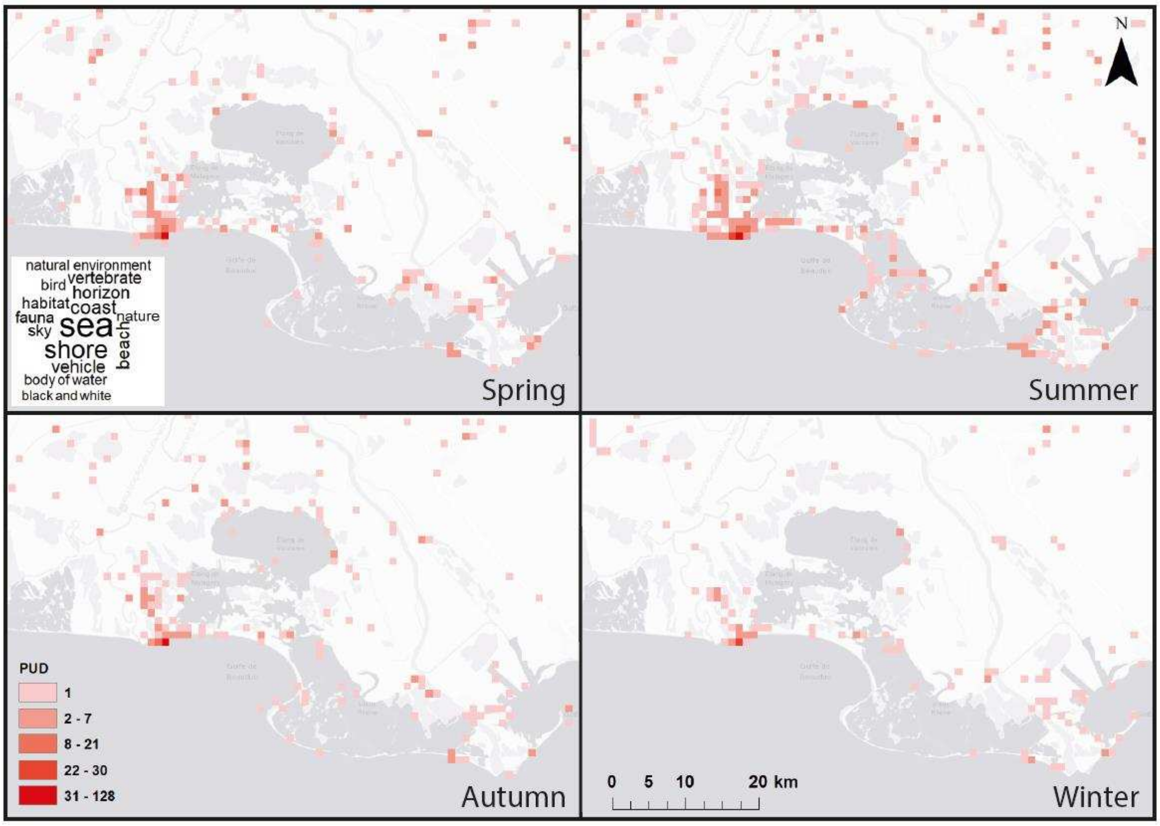

Figure 4: Seasonal distribution of Photo-User Days (PUD) in the Camargue for "nature tourists", with the highest frequency terms shown as a word cloud.

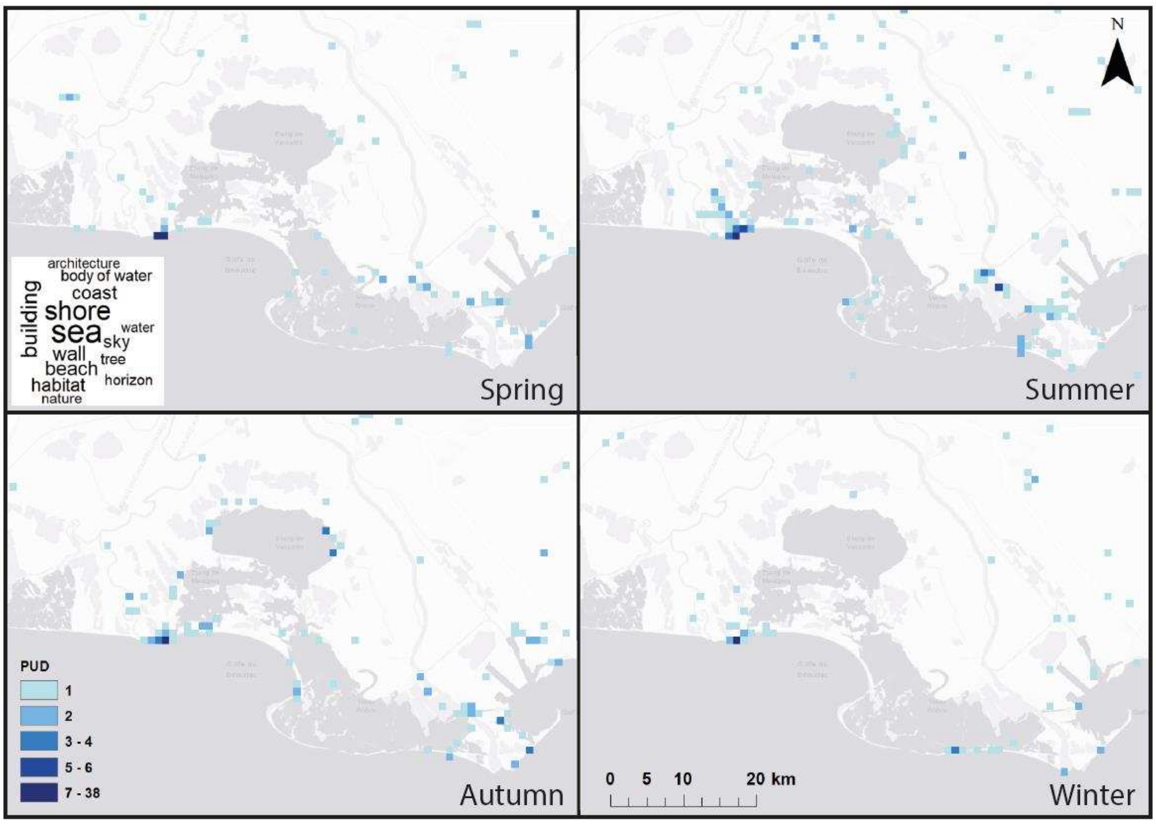

Figure 5: Seasonal distribution of Photo-User Days (PUD) in the Camargue for "general tourists", with the highest frequency terms shown as a word cloud.

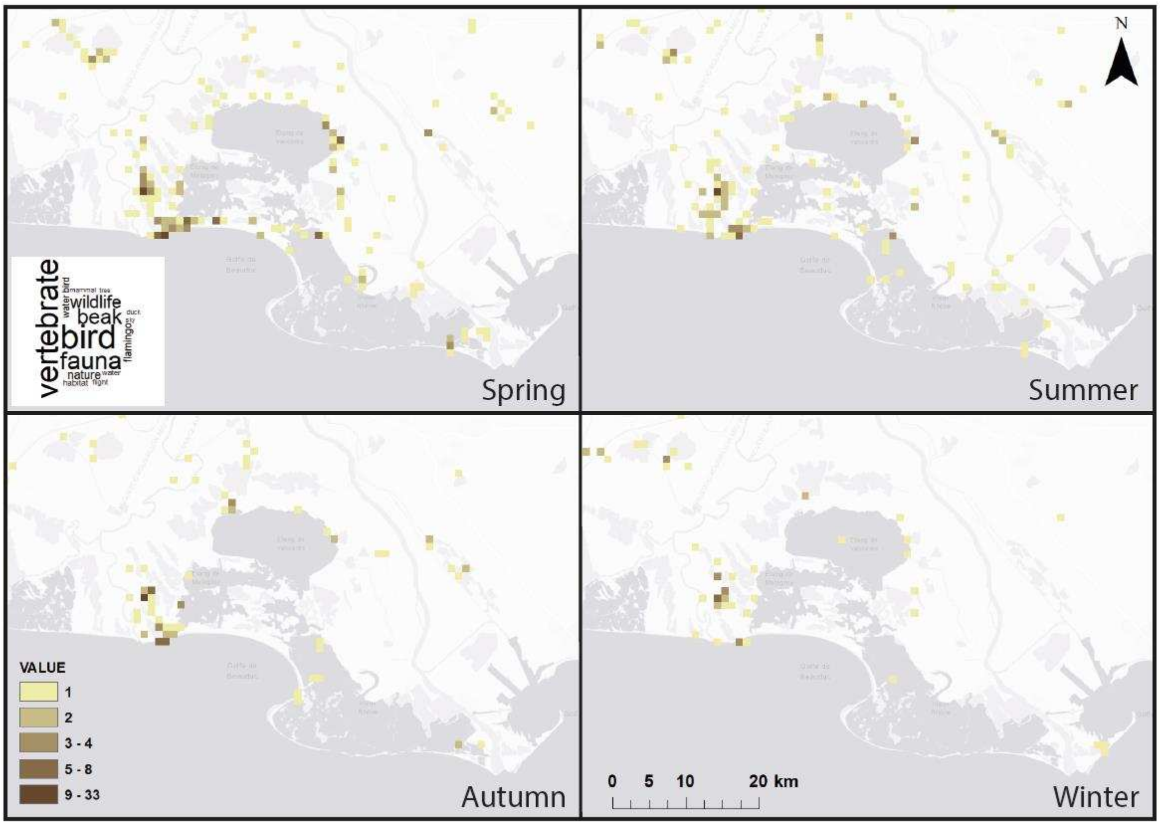

Figure 6: Seasonal distribution of Photo-User Days (PUD) in the Camargue for “bird lovers”, with the highest frequency terms shown as a word cloud.

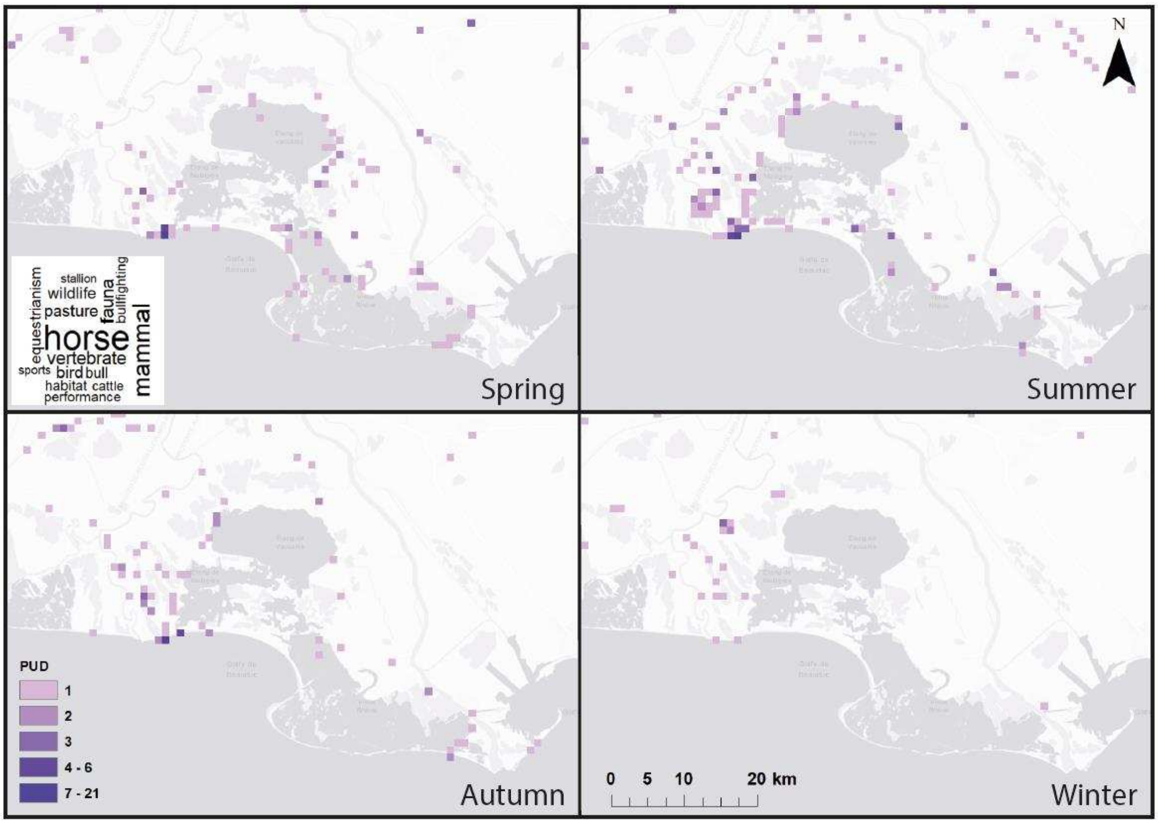

Figure 7: Seasonal distribution of Photo-User Days (PUD) in the Camargue for “equestrian enthusiasts”, with the highest frequency terms shown as a word cloud.

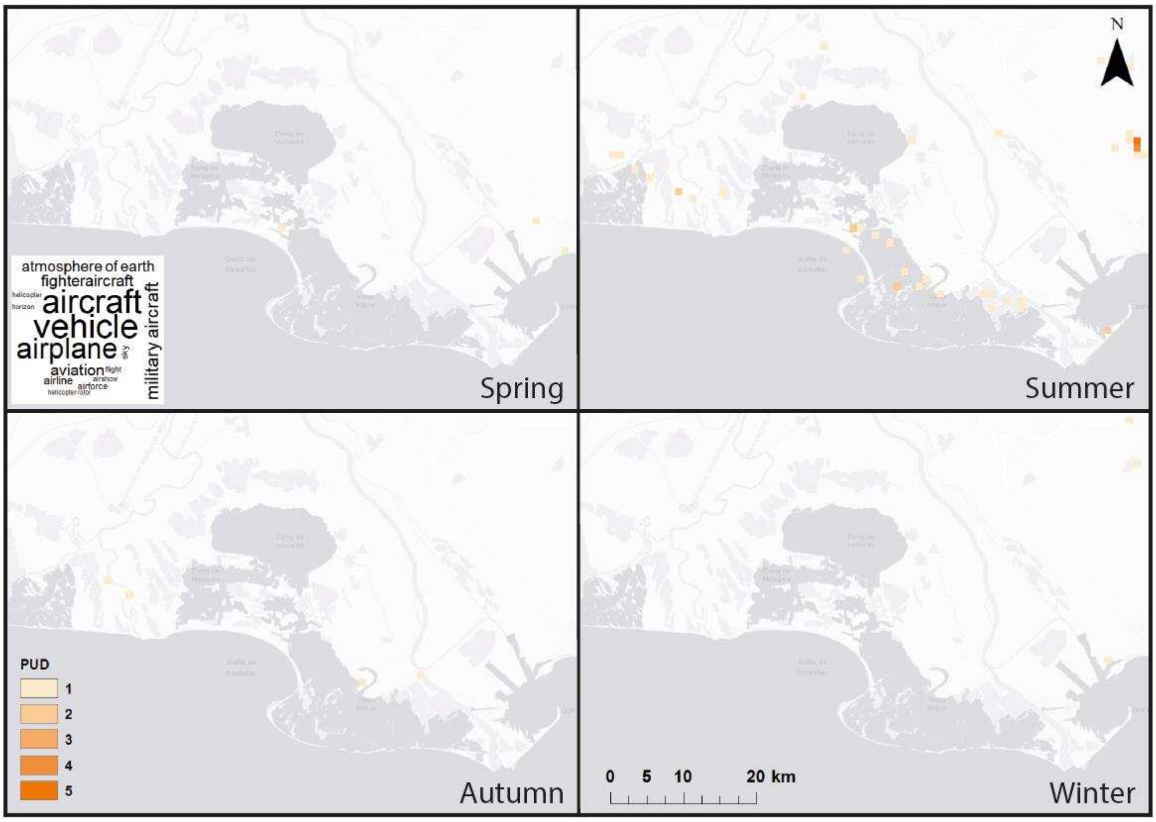

Figure 8: Seasonal distribution of Photo-User Days (PUD) in the Camargue for "aviation enthusiasts", with the highest frequency terms shown as a word cloud.

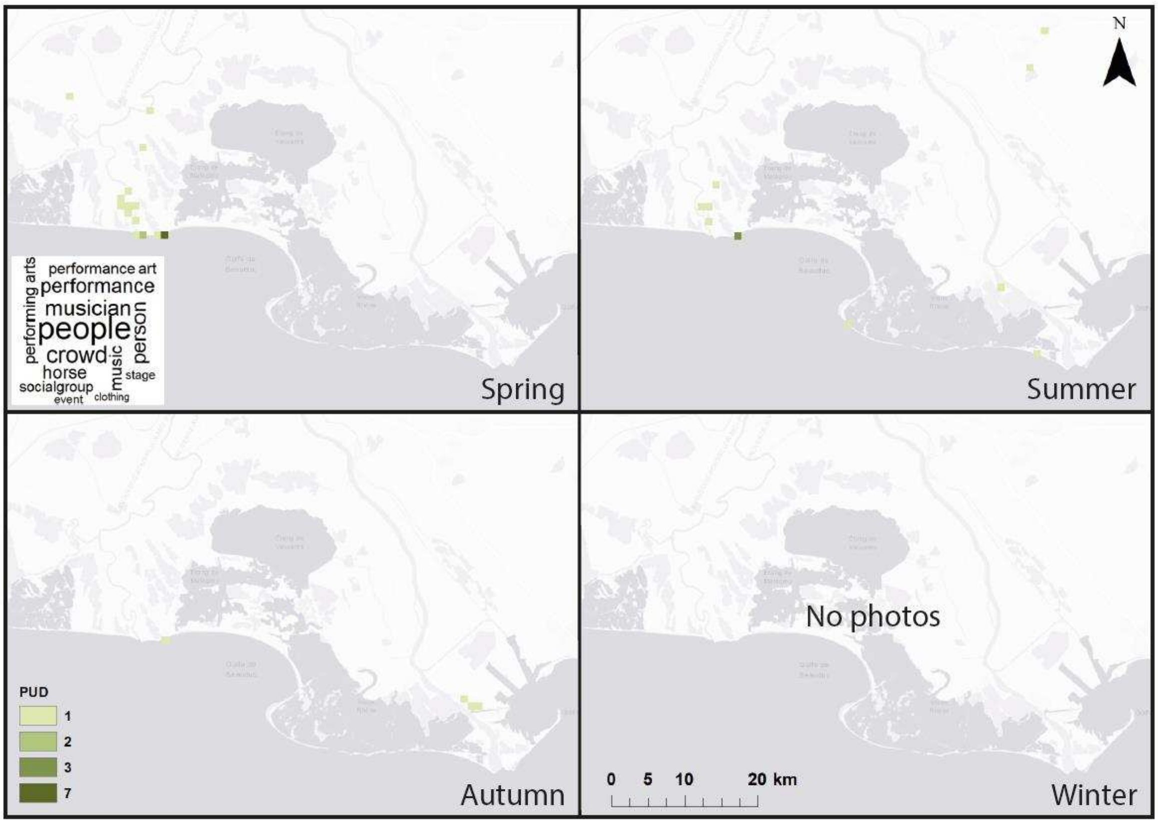

Figure 9: Seasonal distribution of Photo-User Days (PUD) in the Camargue for "religious visitors", with the highest frequency terms shown as a word cloud.

Based on the SOM analysis (Figure 10), five clusters of distinct spatio-temporal patterns of visitation and recreation use were identified across the landscape, with Area 3 covering most of the Camargue with 90.78%. Both Parc Ornithologique de Pont de Gau and Saintes-Maries-de-la-Mer have their own area type (0.13% each), with Area 4 having 3.03% and Area 5 having 5.93% of the pixel cover (Supplemental Information, Table S3). The underlying contribution of each of the six groups to the five SOM clusters can be seen in Table 2.

	Nature tourists	**General tourists**	**Bird lovers**	**Equestrian enthusiasts**	**Aviation enthusiasts**	**Religious visitors**
Area 1	26.71	26.54	9.79	24.40	-0.21	24.88
Area 2	3.53	0.97	24.38	4.76	-0.21	-0.11
Area 3	-0.07	-0.07	-0.05	-0.09	-0.21	-0.11
Area 4	0.57	0.66	0.44	0.78	-0.07	2.29
Area 5	-0.06	-0.11	0.00	-0.10	3.28	-0.06

Table 2: The contributions of the six visitor groups to the SOM identified areas (as z-scores).

Only Area 1, encompassing the village of Saintes-Maries-de-la-Mer is characterised by high PUDs from all groups except aviation enthusiasts, and to a lesser extent bird lovers. From a cultural ES point of view, this area could be considered as a "multifunctional" site. Area 2 for Parc Ornithologique is driven by high PUDs from birdwatchers but is also visited by equestrian enthusiasts and nature tourists. Area 3 gathers all the sites where there were some pictures taken but at very low frequencies. The Area 4 cluster is again characterised by low PUDs in general but is more visited than Area 3; the highest PUDs being related to visitors who come for religious reasons. Area 5 is characterised by high PUDs from aviation enthusiasts who apparently visit the base but also take photos (potentially fly) along the coast.

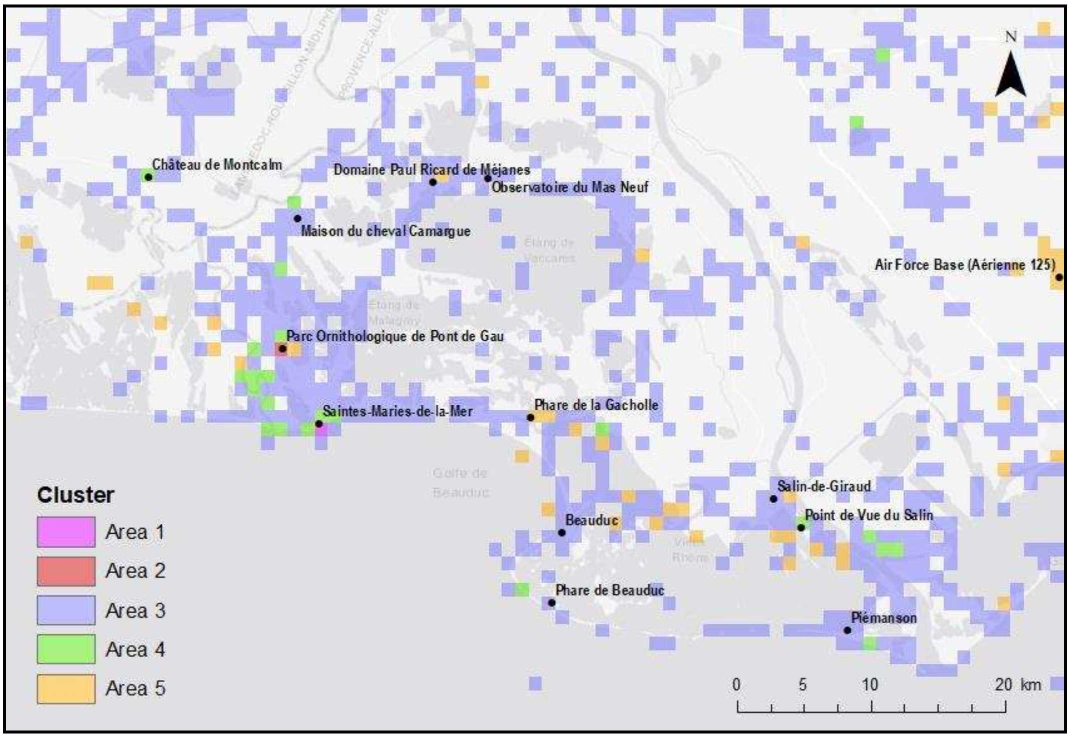

Figure 10: Self-organising map analysis highlighting five clusters of use by different compositions of visitors across the Camargue. Mapping elements: Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community.

Discussion

The importance of identifying beneficiaries is key within the ES framework, and the identification of visitors to create unhomogenised maps of recreation is important for catering to the needs of these visitors. Flickr data analysis demonstrated spatial and temporal visitation patterns of distinct groups of users, information which could contribute to better identification of ES beneficiaries. Using this study approach has two advantages: 1) neutrality in terms of place, groups and seasons and 2) cost- and effort effectiveness. Assessments of visitors often take place in peak seasons (e.g. summer) and at known locations (e.g. the visitor centre) to reach a maximum number of visitors. However, this has implications for the type of visitor that you reach. We found that in the summer the Camargue is predominantly used by birdwatchers and beach visitors while some user groups (e.g. religious our aviation) come more in spring and only at specific locations. The experts were not surprised by these findings, but they were surprised by for instance that the visitors to the music festival in Port St Louis (the village in the east of the Camargue) taking place in the autumn, were grouped in the same category as the pilgrimage to St Marie-de-la-mer (village in the west of the Camargue) which takes place in spring.

Bird and nature are predominant attractions in the Camargue based on surveys of visitors at three sites: Parc Ornithologique de Pont de Gau, Scamandre Centre and Vigueirat Marshes (Chazée *et al.*, 2007). Chazée et al. (2007) found that most visitors could be grouped into 'nature logic' for visiting wetland sites in France, with nature being an important aspect and backdrop of the visit (57%), followed by 'social logic' where meeting friends and family, leisure, visiting tourist places and general social activities are most important (30%). Birdwatching is a fast-growing recreation activity and has been described as a new variant of niche tourism, often attracting affluent tourists (Connell, 2009). Hence identification of these tourists can be beneficial to the local economy, for example, approximately 98 million adults engage in activities such as bird watching, wildlife photography, hunting and fishing spending $59.5 billion on an annual basis in the US alone (Özcan *et al.*, 2009).

Regional attractions are also important for visitors (Chazée et al. 2007). This study identified Saintes-Maries-de-la-Mer and Pont de Gau as being the most important attractions in the Camargue. General tourist A and B groups, based on the word clouds (Figure S3) appreciate the flora and fauna, which is in line with Chazée et al. (2007) who suggested that 87% of those surveyed in the Camargue enjoy and are interested in the observation of fauna (birds and other wildlife) in an aesthetically pleasing and accessible landscape. A more recent study based on participatory mapping showed that wilderness and recreation are the main socio-cultural values attributed to the Camargue landscape based on 113 participants who live or work in the Camargue (Ernoul *et al.*, 2018). While there was strong concurrence between recreational and aesthetic values in coastal zones, areas accessible to the public, beaches and roads surrounding protected areas, it appears that the areas of Saintes-Maries-de-la-Mer and Pont de Gau were not so dominant in the minds of local people as recreational and natural areas.

A study on the Bobrek wetland in Poland found that the local public was divided into two segments regarding management attributes (flood risk, biodiversity and riverbank access [recreation]. A total of 62.5% of users derived positive values for flood risk and riverbank access and a negative value for biodiversity. The remaining users derived positive values for all attributes, though river bank access had the lowest value (Birol *et al.*, 2009)). This contrasts with the present study which infers that most visitors place a positive value on biodiversity, or nature from the types of words that are captured in the word clouds for most groups. The study of Ernoul et al. (2018) suggested that, in contrast to Poland, local people also place a rather positive value on biodiversity in the Camargue.

The case of the aviation enthusiasts and religious visitors and the identification of locations special to these groups in the Camargue infer the methodology is sensitive enough to pick up local differences among group types. The word clouds generated for each group were cohesive and made sense, with several high-frequency terms. Although these groups have small numbers of users, if they were a collection of outliers, then the frequencies of the words would be similar in size in the word cloud diagram, though this is not the case. These groups are small percentages of the Flickr users, though whether they are a small proportion of visitors is a different question, as the aim of this study was to investigate different groups and their spatial patterns, not to quantify the visitor numbers in each group.

Under the SOM analysis, area 3 covered over 90% of all the pixels users visited, showing the impact of low PUDs in the SOM analysis. This demonstrates the need for a minimum number of photos for assessments to provide meaningful results, as, despite moderate numbers of PUDs used for the SOM analysis, we still have a large cluster of low PUD frequency from all groups.

The Flickr analysis allowed to distinguish between different actor groups that are of importance for park managers, however, it also has to be stressed that specific economic sectors and actors were not detected (e.g. farmers, waterboard, heavy transport sector). From the current analysis it is not clear if these groups were not taking/uploading photos or they did not use the recreation ecosystem services, or they did both, but their use of the region cannot be statistically separated from use patterns of the other users. These sectors in the Camargue, and other elements (e.g. age, family composition, origin) could be of importance for park management, but were also not identified. This could be due to biases in the data (elderly do not upload their photos) or due to biases in use of the region (e.g. elderly people do not go into the Camargue). Extracting information from Flickr users' profile may give some information on demographics but was not attempted in this study as all images from Flickr were used and not filtered for the content or user metadata. Not all visitors will take and upload photos onto a social media platform, hence sampling bias is inherent in Flickr and social media data (Levin, Lechner and Brown, 2017; Walden-Schreiner *et al.*, 2018).

Flickr data is biased by factors that are subject to continuous change including the popularity of the platform, user groups and geography (Sessions *et al.*, 2016). Flickr is popular in the US and Western Europe (Levin, Kark and Crandall, 2015), hence was appropriate to use for this study, though it has been found that the demographics of those who post geo-referenced photos online are likely to be well-educated people who work in the fields of arts, science, business or management (Li, Goodchild and Xu, 2013), hence not a representative sample of society. It has been suggested that Flickr users are more likely to share 'high-quality professional photographs' compared to 'every-day experiences' shared by Instagram users, or 'thoughts' by Twitter users, and is the least popular among all three platforms (Tenkanen *et al.*, 2017).

A further limitation for this research was the use of a single photo platform. Though information for the Flickr user base can be found in reports on the internet, the number of Flickr users visiting the Camargue was not available. Hence we cannot remove possible long-term variation in that number which could affect trends in visitation (Figure 2). Geo-tagging errors in photos were identified from an exclusion zone identified during the consultative process with local actors (see Figure S6), though the relatively low numbers did not impact the analysis.

The low average photos taken by nature tourists, general tourists and bird lovers averaged less than 10 images per visitor, compared to over 50 images per religious visitor. This shows how the method allows the spatial distinction between user groups, despite whether they upload little, or large, numbers of photos. This large variation shows that the more niche groups are separated out from the more generalist groups. It could also mean that users uploading more images of the same content could influence the final groups; though it must be noted that PUD was used, hence these images are over a broader range of 1 km pixels and days. Hence the users are also more intensive or high-frequency visitors to the areas. Additionally, without an extensive network of known visitation numbers for various parts of the landscape, a regression to convert PUD to visitors cannot be robustly undertaken.

Other potential weaknesses in the methodology are the image annotation and LSA. Google Cloud Vision has been used by several studies to analyse the content of images (Hyam, 2017; Richards and Tunçer, 2017) though may miss or mislabel content, for example, subjective assessment by Hyam (2017) found that the natural subject missed was high, though false positives were low. The use of LSA has several disadvantages including being computationally expensive and difficult to implement for the practitioner, with defining the number of dimensions for the matrix being a 'balancing act' between capturing latent semantic information and reducing noise (Miller, 2003). For future expansion on this research, the role of biotic and abiotic factors could be assessed, with the inclusion of remotely sensed data to monitor the impact of seasonal events and larger temporal events, such as temporal ponding on the different visitor groups. Additionally, we could separate users by place of origin, hence be able to distinguish between recreation or tourism or local/domestic and foreign visitors as demonstrated by Juutinen et al. (2011) to investigate the differences in ecological and recreation preferences in Oulanka National Park in Finland. As this paper does not distinguish between the types of photos taken, future research could also filter for indoor/outdoor photos with the filtering of Google Cloud Vision image annotations or photo metadata directly from Flickr.

It is clear that park managers will very likely not be able to use raw social media data themselves directly and would need a tool developed to facilitate user-friendly harvesting and interpretation of data, but once in place, this could be a much more effort and cost-effective method than doing surveys in the field. This study has identified information which has been received by managers in the Camargue as very interesting. In particular, knowing when and where bird watchers and nature lovers wander in the Camargue is considered as original knowledge because these tourists often go undetected while touring in the Camargue. Using the obtained maps, we asked the representatives of the park management whether and how they would use the obtained information. They indicated that the maps confirmed important assumptions on tourism in the area, such as the limited use that religious and beach tourists make of the wider Camargue region. Having a closer look at the pictures taken by these people could help park managers to develop a more strategic and efficient promotion of other areas likely to be appreciated by these visitors. When asking targeted questions, several potential uses could be identified by park managers — for instance, using the maps to identify locations for specific user groups or to seek potential collaborations to promote awareness of natural richness (e.g. the horse museum). Campaigns could then be targeted at user groups and/or at specific periods to increase recreational activity in some areas and decreasing it in others.

Conclusion

By obtaining a quantification of the use of the Camargue, arguments can be developed to influence regional decisions. For instance, on the maintenance of roads or the construction of barriers to either improve or reduce accessibility. An understanding of visitor types in similar protected areas can guide the development of sustainable ecotourism in other areas. Globally the recreation and tourism industry is economically significant, contributing to many regional economies (Wood *et al.*, 2013). The growing trend in nature-based recreation (Balmford *et al.*, 2009) highlights the need for areas that match visitors needs in recreational areas. Studies have quantified that factors such as temperature, precipitation, infrastructure and habitat diversity and species richness are important in varying degrees for recreation for visitors (Jones and Scott, 2006; Neuvonen *et al.*, 2010; Juutinen *et al.*, 2011; Wood *et al.*, 2013; Siikamäki *et al.*, 2015; Millhäusler *et al.*, 2016). The utilisation of techniques that allow different and/or unique beneficiary groups to be analysed separately will allow more nuanced and dynamic management strategies to be developed for recreational areas.

Social media data can be harnessed to better understand the area where visitors place value. Geo-referenced images coupled with content analysis allow a greater understanding of not only where users visit, but what especially they find attractive in the environment. By harnessing the power of LSA in this study, we have been able to demonstrate how visitors can be grouped to visualise spatial and temporal patterns of visitation. With increasing pressure on protected areas, this type of analysis can allow park managers and decision makers to see how proposed management may impact respective beneficiary groups.

Funding: This work was supported by ECOPOTENTIAL under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 641726). This research is a contribution to the GEO BON working group on ecosystem services.

References

Arnberger, A. and Eder, R. (2011) 'The influence of age on recreational trail preferences of urban green-space visitors: a discrete choice experiment with digitally calibrated images', *Journal of Environmental Planning and Management*, 54(7), pp. 891–908. doi: 10.1080/09640568.2010.539875.

Bagstad, K. J. *et al.* (2014) 'From theoretical to actual ecosystem services: Mapping beneficiaries and spatial flows in ecosystem service assessments', *Ecology and Society*, 19(2). doi: 10.5751/ES-06523-190264.

Balmford, A. *et al.* (2009) 'A global perspective on trends in nature-based tourism', *PLoS Biology*, 7(6), pp. 1–6. doi: 10.1371/journal.pbio.1000144.

Van Berkel, D. B. *et al.* (2018) 'Quantifying the visual-sensory landscape qualities that contribute to cultural ecosystem services using social media and LiDAR', *Ecosystem Services*. doi: 10.1016/j.ecoser.2018.03.022.

Birol, E. *et al.* (2009) 'Optimal management of wetlands: Quantifying trade-offs between flood risks, recreation, and biodiversity conservation', *Water Resources Research*. Wiley-Blackwell, 45(11). doi: 10.1029/2008WR006955.

Blondel, J., G. Barruol, and R. Vianet. (2013). *L'Encyclopédie de la Camargue*. Buchet-Chastel, Paris, France.

Boxall, P. C. and Adamowicz, W. L. (2002) 'Understanding Heterogeneous Preferences in Random Utility Models : A Latent Class Approach', *Environmental Resoure Econonmics*, 23(4), pp.421–446.

Charrad, M. *et al.* (2014) '{NbClust}: An {R} Package for Determining the Relevant Number of Clusters in a Data Set', *Journal of Statistical Software*, 61(6), pp. 1–36. Available at: http://www.jstatsoft.org/v61/i06/.

Chazée, L. *et al.* (2007) *Recreational and educational services of Mediterranean Wetlands*. Arles, France.

Cheng, X. *et al.* (2019) 'Evaluation of cultural ecosystem services: A review of methods', *Ecosystem Services*. Elsevier B.V., 37(April), p. 100925. doi: 10.1016/j.ecoser.2019.100925.

Connell, J. (2009) 'Birdwatching, twitching and tourism: Towards an Australian perspective', *Australian Geographer*. Taylor & Francis Group, 40(2), pp. 203–217. doi: 10.1080/00049180902964942.

Ehrlich, O. *et al.* (2017) 'A latent class analysis of public attitudes toward water resources with implications for recreational demand', *Ecosystem Services*. Elsevier B.V., 28, pp. 124–132. doi: 10.1016/j.ecoser.2017.10.019.

Ernoul, L. *et al.* (2018) 'Participatory mapping: Exploring landscape values associated with an iconic species', *Applied Geography*. Elsevier Ltd, 95(April), pp. 71–78. doi: 10.1016/j.apgeog.2018.04.013.

Feld, C. K. *et al.* (2009) 'Indicators of biodiversity and ecosystem services: A synthesis across ecosystems and spatial scales', *Oikos*, 118(12), pp. 1862–1871. doi: 10.1111/j.1600-0706.2009.17860.x.

García-Nieto, A. P. *et al.* (2015) 'Collaborative mapping of ecosystem services: The role of stakeholders' profiles', *Ecosystem Services*, 13, pp. 141–152. doi: 10.1016/j.ecoser.2014.11.006.

Geijzendorffer, I. R. *et al.* (2017) ‘Ecosystem services in global sustainability policies’, *Environmental Science & Policy*, 74(January), pp. 40–48. doi: 10.1016/j.envsci.2017.04.017.

Gentin, S. (2011) ‘Outdoor recreation and ethnicity in Europe-A review’, *Urban Forestry and Urban Greening*, 10(3), pp. 153–161. doi: 10.1016/j.ufug.2011.05.002.

Google Cloud Vision (2017) *Documentation for the Google Cloud Vision API*. Available at: www.cloud.google.com/vision/.

Gosal, A. S., Newton, A. C. and Gillingham, P. K. (2018) ‘Comparison of methods for a landscape-scale assessment of the cultural ecosystem services associated with different habitats’, *International Journal of Biodiversity Science, Ecosystem Services & Management*. Taylor & Francis, 14(1), pp. 91–104. doi: 10.1080/21513732.2018.1447016.

Guerrero, P. *et al.* (2016) ‘Revealing Cultural Ecosystem Services through Instagram Images: The Potential of Social Media Volunteered Geographic Information for Urban Green Infrastructure Planning and Governance’, *Urban Planning*, 1(2), p. 1. doi: 10.17645/up.v1i2.609.

Hadwen, W. L., Hill, W. and Pickering, C. M. (2008) ‘Linking visitor impact research to visitor impact monitoring in protected areas’, *Journal of Ecotourism*, 7(1), pp. 87–93. doi: 10.2167/joe193.0.

Haines-Young, R. and Potschin, M. (2012) ‘Common international classification of ecosystem services (CICES, Version 4.1)’, *European Environment Agency*, 33.

Hausmann, A. *et al.* (2017) ‘Social Media Data Can Be Used to Understand Tourists’ Preferences for Nature-Based Experiences in Protected Areas’, *Conservation Letters*, 11(1), p. n/a-n/a. doi: 10.1111/conl.12343.

Heath, M. F. *et al.* (2000) ‘Important Bird Areas in Europe: priority sites for conservation’.

Hermes, J. *et al.* (2018) ‘Assessment and valuation of recreational ecosystem services of landscapes’, *Ecosystem Services*. Elsevier B.V., 31, pp. 289–295. doi: 10.1016/j.ecoser.2018.04.011.

Hernández-Morcillo, M., Plieninger, T. and Bieling, C. (2013) ‘An empirical review of cultural ecosystem service indicators’, *Ecological Indicators*. Elsevier Ltd, 29(October 2017), pp. 434–444. doi: 10.1016/j.ecolind.2013.01.013.

Hyam, R. (2017) ‘Automated Image Sampling and Classification Can Be Used to Explore Perceived Naturalness of Urban Spaces’, *PLoS One*, 12(1), p. e0169357.

Jones, B. and Scott, D. (2006) ‘Climate Change, Seasonality and Visitation to Canada’s National Parks.’, *Journal of Park & Recreation Administration*, 24(2), pp. 42–62. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=21792416&site=ehost-live.

Juutinen, A. *et al.* (2011) ‘Combining ecological and recreational aspects in national park management: A choice experiment application’, *Ecological Economics*. Elsevier B.V., 70(6), pp. 1231–1239. doi: 10.1016/j.ecolecon.2011.02.006.

Kitchin, R. (2014) ‘Big Data, new epistemologies and paradigm shifts’, *Big Data & Society*, 1(1), p. 2053951714528481.

Kutz, J. N. *et al.* (2016) *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM.

Landauer, T. K., Foltz, P. W. and Laham, D. (1998) ‘An introduction to latent semantic analysis’, *Discourse processes*, 25(2–3), pp. 259–284.

Levin, N., Kark, S. and Crandall, D. (2015) ‘Where have all the people gone? Enhancing global conservation using night lights and social media', *Ecological Applications*, 25(8), pp. 2153–2167. doi: 10.1890/15-0113.1.

Levin, N., Lechner, A. M. and Brown, G. (2017) 'An evaluation of crowdsourced information for assessing the visitation and perceived importance of protected areas', *Applied Geography*. Elsevier Ltd, 79, pp. 115–126. doi: 10.1016/j.apgeog.2016.12.009.

Li, L., Goodchild, M. F. and Xu, B. (2013) 'Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr', *Cartography and Geographic Information Science*, 40(2), pp. 61–77. doi: 10.1080/15230406.2013.777139.

Martin-Lopez, B. *et al.* (2012) 'Uncovering ecosystem service bundles through social preferences', *PLoS ONE*, 7(6). doi: 10.1371/journal.pone.0038970.

Martin-Ortega, J. *et al.* (2019) 'Nature commodification: "a necessary evil"? An analysis of the views of environmental professionals on ecosystem services-based approaches', *Ecosystem Services*. Elsevier B.V., 37(April), p. 100926. doi: 10.1016/j.ecoser.2019.100926.

Milcu, A. *et al.* (2013) 'Cultural Ecosystem Services: A Literature Review and Prospects for Future Research', *Ecology and Society*, 18(3), p. 44. doi: 10.5751/ES-05790-180344.

Miller, T. (2003) 'Essay assessment with latent semantic analysis', *Journal of Educational Computing Research*, 29(4), pp. 495–512. doi: 10.2190/W5AR-DYPW-40KX-FL99.

Millhäusler, A. *et al.* (2016) 'Publicity, economics and weather – Changes in visitor numbers to a European National Park over 8 years', *Journal of Outdoor Recreation and Tourism*. Elsevier, 16(September), pp. 50–57. doi: 10.1016/j.jort.2016.09.005.

Di Minin, E., Tenkanen, H. and Toivonen, T. (2015) 'Prospects and challenges for social media data in conservation science', *Frontiers in Environmental Science*, 3(September), p. 63. doi: 10.3389/fenvs.2015.00063.

Nahlik, A. M. *et al.* (2012) 'Where is the consensus? A proposed foundation for moving ecosystem service concepts into practice', *Ecological Economics*. Elsevier B.V., 77, pp. 27–35. doi: 10.1016/j.ecolecon.2012.01.001.

Neuvonen, M. *et al.* (2010) 'Visits to national parks: Effects of park characteristics and spatial demand', *Journal for Nature Conservation*, 18(3), pp. 224–229. doi: 10.1016/j.jnc.2009.10.003.

Oteros-Rozas, E. *et al.* (2017) 'Using social media photos to explore the relation between cultural ecosystem services and landscape features across five European sites', *Ecological Indicators*. Elsevier Ltd. doi: 10.1016/j.ecolind.2017.02.009.

Özcan, H. *et al.* (2009) 'Ecotourism Potential and Management of Kavak Delta (Northwest Turkey)', *Journal of Coastal Research*. Coastal Education and Research Foundation, 253, pp. 781–787. doi: 10.2112/08-1068.1.

R Core Team (2017) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria. Available at: https://www.r-project.org/.

Richards, D. R. and Friess, D. A. (2015) 'A rapid indicator of cultural ecosystem service usage at a fine spatial scale: Content analysis of social media photographs', *Ecological Indicators*. Elsevier Ltd, 53, pp. 187–195. doi: https://doi.org/10.1016/j.ecolind.2015.01.034.

Richards, D. R. and Tunçer, B. (2017) 'Using image recognition to automate assessment of cultural ecosystem services from social media photographs', *Ecosystem Services*. Elsevier B.V. doi: 10.1016/j.ecoser.2017.09.004.

Richards, D. R. and Tunçer, B. (2018) 'Using image recognition to automate assessment of cultural ecosystem services from social media photographs', *Ecosystem Services*. Elsevier B.V., 31, pp. 318–325. doi: 10.1016/j.ecoser.2017.09.004.

Scarpa, R. and Thiene, M. (2004) 'Destination Choice Models for Rock Climbing in the Northeastern Alps : A Latent-Class Approach Based on Intensity of Preferences', (June).

Seresinhe, C. I., Preis, T. and Moat, H. S. (2017) 'Using deep learning to quantify the beauty of outdoor places', *Royal Society Open Science*, 4(7), p. 170170. doi: 10.1098/rsos.170170.

Sessions, C. *et al.* (2016) 'Measuring recreational visitation at U.S. National Parks with crowd-sourced photographs', *Journal of Environmental Management*. Elsevier Ltd, 183, pp. 703–711. doi: 10.1016/j.jenvman.2016.09.018.

Siikamäki, P. *et al.* (2015) 'Biodiversity attracts visitors to national parks', *Biodiversity and Conservation*, 24(10), pp. 2521–2534. doi: 10.1007/s10531-015-0941-5.

Sonter, L. J. *et al.* (2016) 'Spatial and temporal dynamics and value of nature-based recreation, estimated via social media', *PLoS ONE*, 11(9), pp. 1–16. doi: 10.1371/journal.pone.0162372.

Tenerelli, P., Demšar, U. and Luque, S. (2016) 'Crowdsourcing indicators for cultural ecosystem services: A geographically weighted approach for mountain landscapes', *Ecological Indicators*, 64, pp. 237–248. doi: 10.1016/j.ecolind.2015.12.042.

Tenkanen, H. *et al.* (2017) 'Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas', *Scientific Reports*. Springer US, 7(1), pp. 1–11. doi: 10.1038/s41598-017-18007-4.

Walden-Schreiner, C. *et al.* (2018) 'Using crowd-sourced photos to assess seasonal patterns of visitor use in mountain-protected areas', *Ambio*. Springer Netherlands. doi: 10.1007/s13280-018-1020-4.

Walden-Schreiner, C., Leung, Y. F. and Tateosian, L. (2018) 'Digital footprints: Incorporating crowdsourced geographic information for protected area management', *Applied Geography*. Elsevier Ltd, 90(November 2017), pp. 44–54. doi: 10.1016/j.apgeog.2017.11.004.

Wehrens, R. and Buydens, L. M. C. (2007) 'Self- and Super-organising Maps in R: the kohonen package', *J. Stat. Softw.*, 21(5). Available at: http://www.jstatsoft.org/v21/i05.

Wild, F. (2015) 'lsa: Latent Semantic Analysis'. Available at: https://cran.r-project.org/package=lsa.

Wood, S. A. *et al.* (2013) 'Using social media to quantify nature-based tourism and recreation.', *Scientific reports*. Nature Publishing Group, 3(1), p. 2976. doi: 10.1038/srep02976.

Supplementary Information

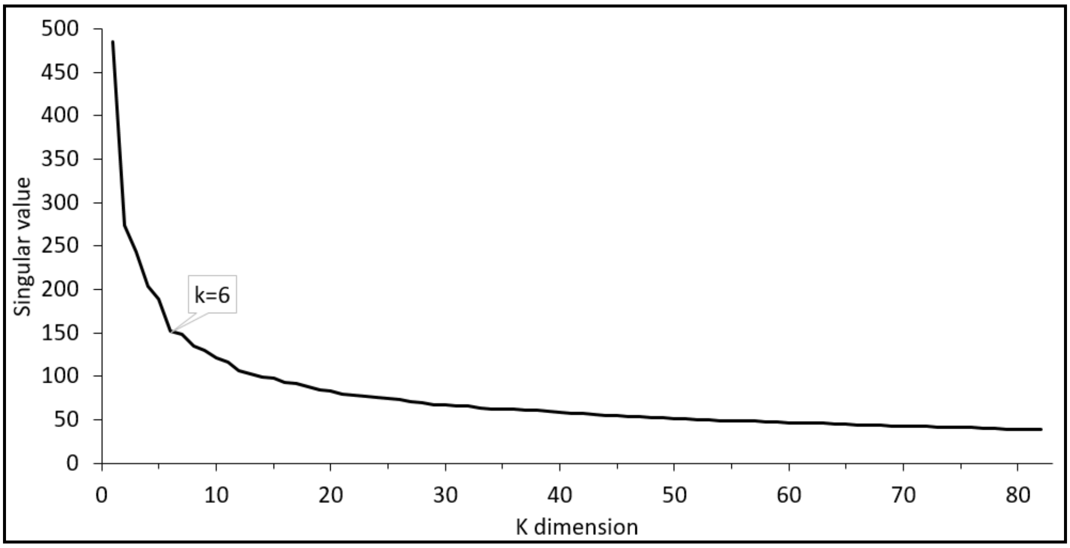

Figure S1: Graph illustrating the variance of the full SVD decomposition, with a total of 82 dimensions. The elbow of the plot is highlighted at $k_{SVD} = 6$. The *lsa* r package (Wild, 2015) provides truncated matrices T_k, S_k and D_k.

Clusters	**Recommended by number of indices**
2	4
3	1
4	1
5	1
6	**6**
7	2
12	3
14	3
15	2

Table S1: Table showing from a total of 23 indices implemented in nbClust for the data, the majority (6) recommended 6 clusters with criteria for cluster selection and the index value. A range of 2-15 clusters was chosen for the analysis. No recommendations were made for between 8-11, and 13 clusters by any index (these have thus been removed from the table).

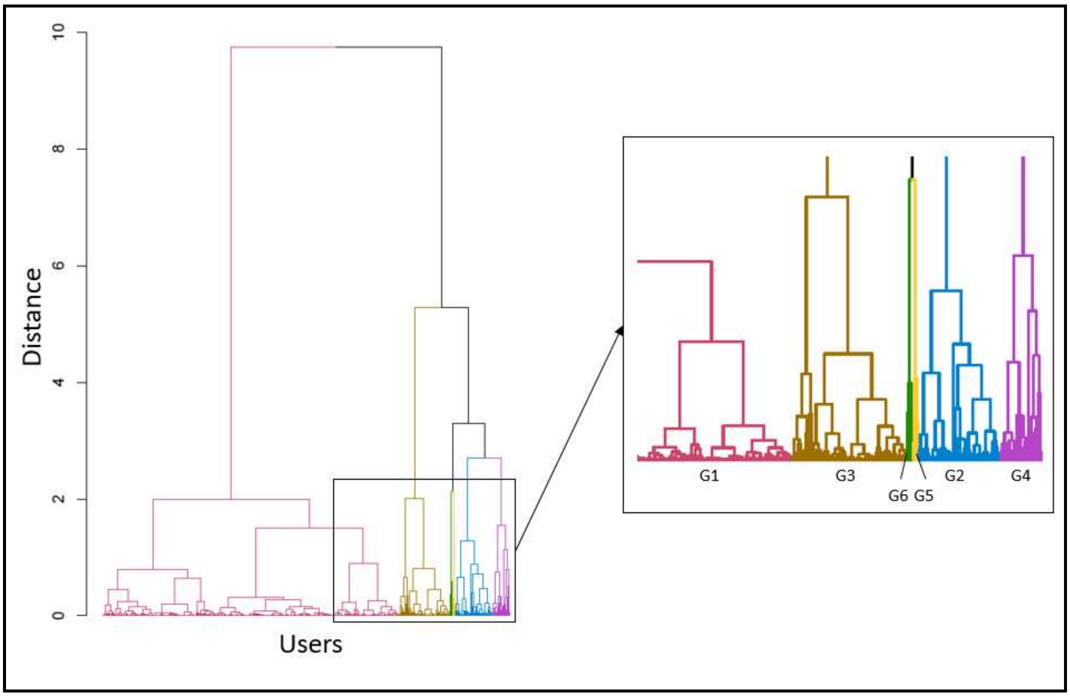

Figure S2: Dendrogram illustrating six groups of users identified from a majority of clustering indices.

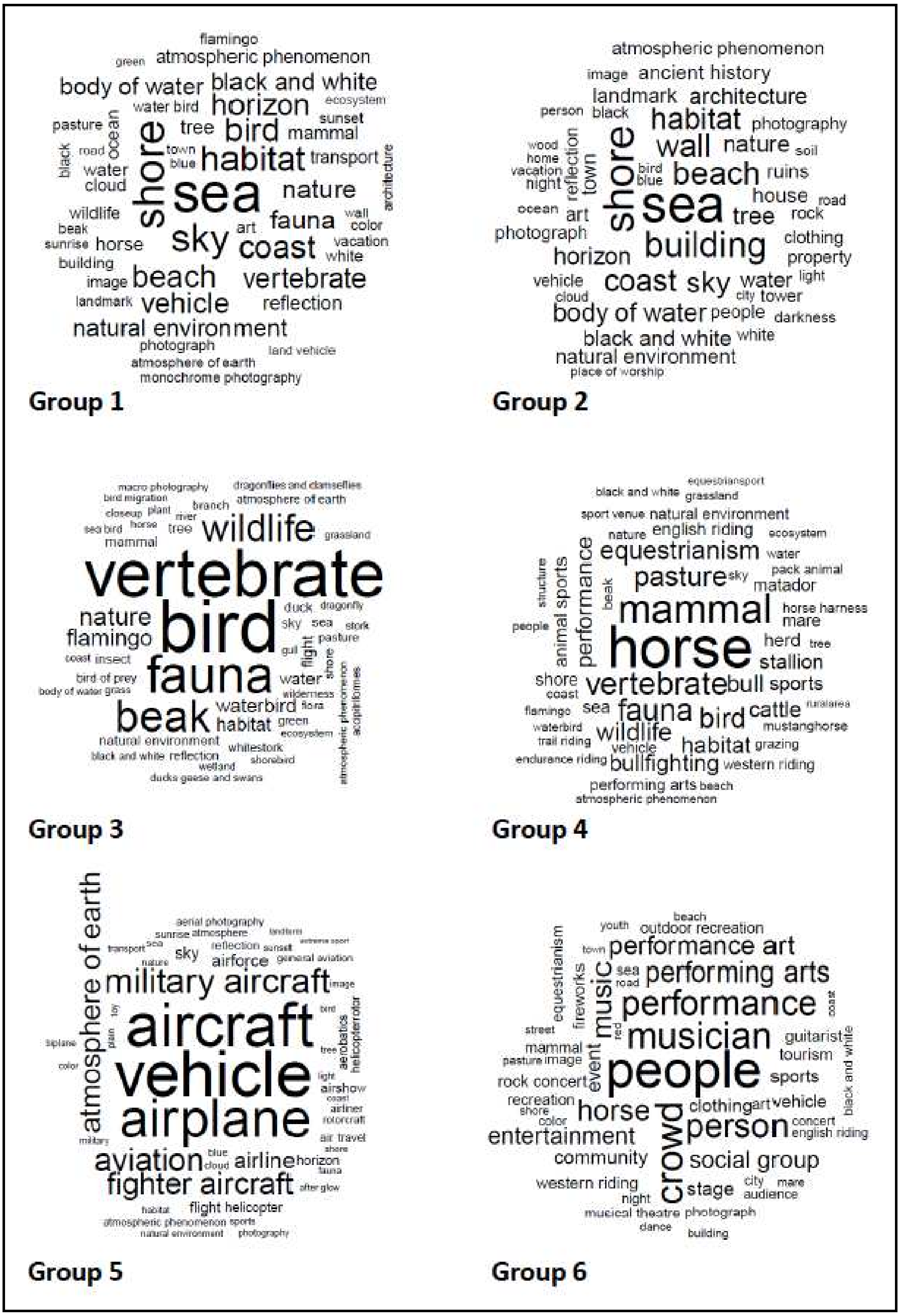

Figure S3: Wordclouds illustrating the top 50 words for the six identified groups.

Group	**S**	**P value**
1	54	0.018
2	37	0.103
3	94	0.000
4	75	0.001
5	16	0.327
6	20	0.253
Ungrouped	97	0.000

Table S2: Results from Seasonal Mann-Kendall trend test on 2007 – 2016 PUD data using ‘trend’ *r* package (Pohlert, 2018).

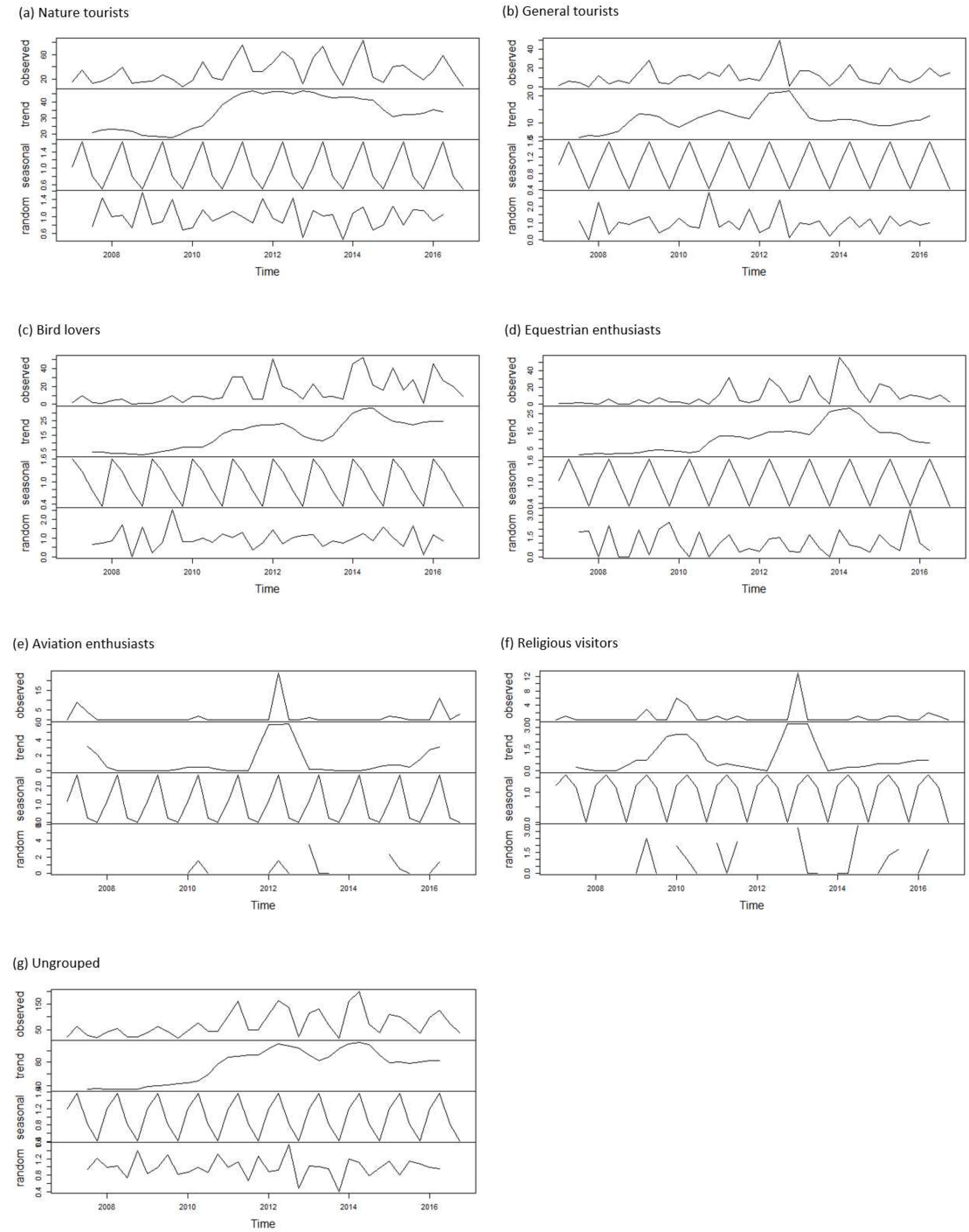

Figure S4: Data was decomposed as under an additive model in R.

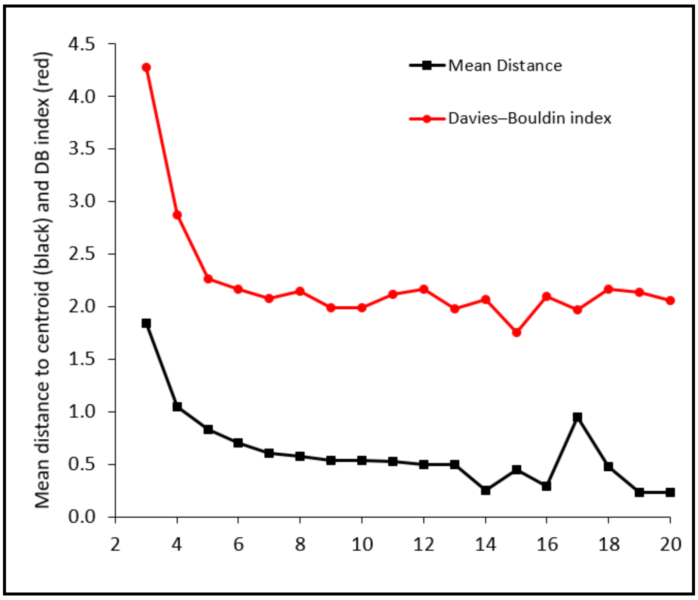

Figure S5: Davies-Bouldin Index and mean distance plot. Five clusters were chosen as an optimum number from the stabilisation seen in the DB index and the moderately low value of the mean SOM distance.

Name	**Area (km^2)**	**Percentage of total area**
Area 1	1	0.13
Area 2	1	0.13
Area 3	719	90.78
Area 4	24	3.03
Area 5	47	5.93

Table S3: SOM identified clusters, with corresponding km^2 areas (relating to each 1 x 1 km pixel).

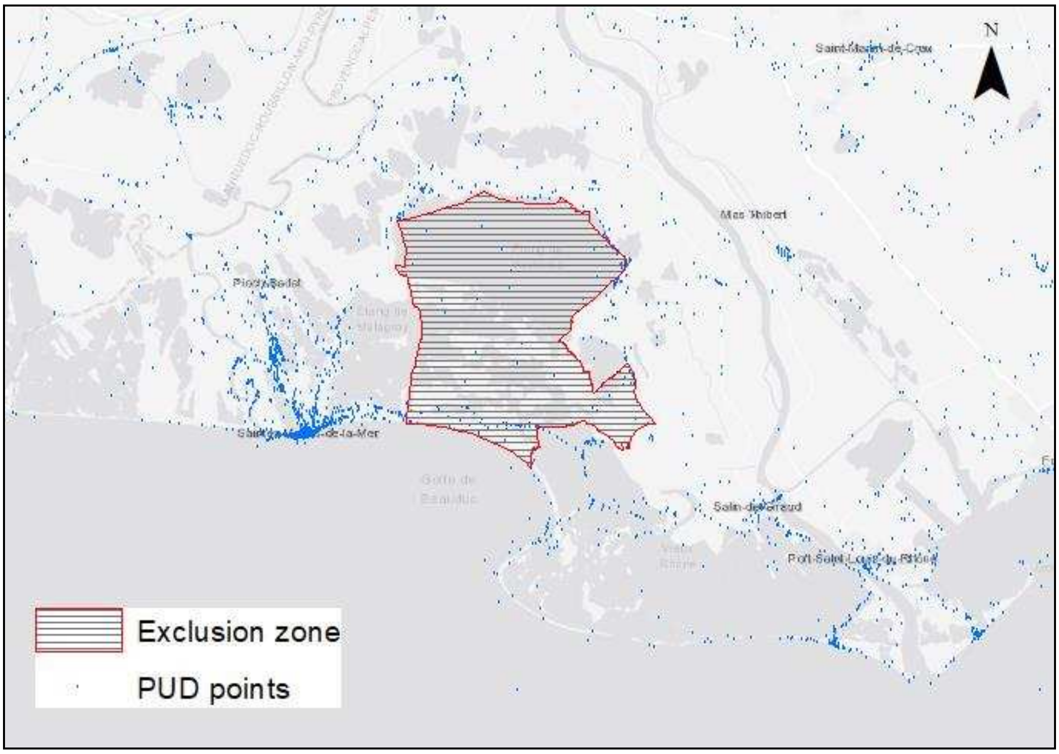

Figure: S6: The zone where no visitors are allowed to enter in the Camargue is highlighted, with 3.26% of total PUD points used within this study situated within the zone, showing that the photo self-geotagged by Flickr users can introduce some error. (Note: Hiking and horse riding is allowed at the southernmost part of the exclusion zone along the beach, further details can be found at http://www.snpn.com/reservedecamargue/). Source: Esri, HERE, DeLorme, MapmyIndia and © OpenStreetMap contributors and the GIS community. Exclusion zone shapefile: Tour du Valat.

References

Pohlert, T., 2018. trend: Non-Parametric Trend Tests and Change-Point Detection. Available at: https://cran.r-project.org/package=trend.

Wild, F., 2015. lsa: Latent Semantic Analysis. Available at: https://cran.r-project.org/package=lsa.