OPEN DATA CUBE APPLICATION TO USER-GENERATED GEODATA: VISITORS TURNOUT INVESTIGATION IN THE INSUBRIA NATURAL PARKS

D. Oxoli[1]*, A. Vavassori[1], J. R. Cedeno Jimenez[1], M. A. Brovelli[1]

[1] Dept. of Civil and Environmental Engineering, Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milano, Italy - (daniele.oxoli, alberto.vavassori, jesusrodrigo.cedeno, maria.brovelli)@polimi.it

Commission IV, WG IV/4

KEY WORDS: Open Data Cube, User-generated Geodata, Natural Parks, Tourism Management

ABSTRACT:

Green areas such as natural and periurban parks embed key assets for biodiversity and landscape preservation and environmental education. In most cases, they also boost local economic growth for their hosting territories thanks to the establishment of eco-tourism activities. Nevertheless, coordinated management and promotion actions between policymakers and promoters remain vital for the sustainable exploitation of these green areas. The knowledge of visitors dynamics across natural parks is necessary information necessary to drive parks management policies while being often a complex data to collect. To that end, this paper proposes the use of user-generated geodata, namely users' location records provided by the Facebook Data for Good program, to assess visitors' turnout for the natural parks of the Insubria Region, between Southern Switzerland and Northern Italy. The Open Data Cube technology, originally developed for managing satellite Earth observation data, was here adapted for processing and analysing the considered user-generated geodata. Space-time patterns of Facebook users' presence in the period May 2020 - December 2021 were extracted to infer visitors' fluxes and destinations preferences. Results pointed out differences between Italian and Swiss parks by outlining also most visited locations within each park. Despite limitations related to data representativeness (limited sample of Facebook app users with location history enabled on their devices) the integration of user-generated geodata in a cutting-edge and free and open-source data management platform, such as the Open Data Cube, turned out to be promising for the improvement of natural areas management practices.

1. INTRODUCTION

The conservation and valorisation of green areas such as natural and periurban parks are critical tasks to support biological diversity, landscape, and environmental education (Markevych et al., 2017) as well as to unpin their tourism potential, which plays a pivotal role in the economy of the hosting territories by sustaining local growth through the establishment of eco-tourism industries (Di Minin et al., 2015). The physical presence of such areas does not directly ensure the above benefits are achieved. Coordinated planning, management and promotion actions between policymakers and promoters are key for ensuring quality, attractiveness and a reduced environmental impact for eco-tourism destinations and activities (Wood, 2002). Nevertheless, these tasks may be hindered by a lack of resources and by fragmented local political contexts to which these areas may be subjected. This can prevent optimal management due to competitive actions as well as weak cooperation among the policymakers and promoters (Prell et al., 2016).

This is the case, for example, of the natural parks of the Insubria Region, a historical-geographical region stretching between the Canton Ticino (Southern Switzerland) and the Lombardy Region (Northern Italy) (Oxoli et al., 2019), which constitute green corridors connecting the Padana Plain to the Alpine environment by crossing many densely populated urban centres (see Figure 1a and 1b).

Among the primary information needs that are necessary for evidence-based tourism management operations in natural parks, there is the monitoring and understanding of visitors turnout and destination preferences that can be used by parks managers to shape local promotion strategies (Hausmann et al., 2018). The investigation of visitors' fluxes within natural parks demands disaggregated and space-time resolved data. The collection of such data by means of traditional surveys and interviews may be resource-demanding and -in many cases- noneffective, to achieve sufficient space-time coverage for extensive analyses in wide natural areas. To that end, social media, crowdsourcing, and citizen-generated geodata have proved to be valuable data sources for the study of visitors' dynamics thanks to their unequalled availability and granularity (Wood et al., 2013).

In view of the above, this study investigates the use of user-generated geodata, namely geolocated time series of users' counts provided by the Facebook Data for Good initiative (Facebook, 2021), to support the analysis and monitoring of people fluxes and to derive metrics on parks recreation and tourism values through comparisons with surrounding areas. This topic has been preliminarily investigated by the authors (Oxoli and Brovelli, 2021). The goal of the present study is to enrich and complete the above preliminary results by introducing cutting-edge data management software tools in the analysis procedures, namely the Open Data Cube (ODC) (Killough, 2018), as a supporting tool for user-generated geodata management. Despite the ODC has been originally developed to support the exploitation of satellite Earth Observation data, here its applicability to not conventional data sources, such as the user-generated geodata from Facebook, is tested and discussed. The geolocated time series of Facebook users' counts processed with the ODC were finally used to investigate visit-

* Corresponding author

ors turnout dynamics within the Insubria Region natural parks. An indicator for the identification of visitors destination preferences (Vavassori et al., 2022) was computed and mapped starting from the raw data processed in the ODC. Results allowed discerning space-time visitors' turnout patterns characterizing each park area by spotting differences with the surrounding areas.

The remainder of this paper is as follows. Section 2 describes the ODC application to the Facebook datasets adopted in this work. Section 3 reports details on data processing and achieved outcomes for the investigation of visitors turnout in the Insubria parks. Finally, conclusions and future directions for the work are presented in Section 4.

2. OPEN DATA CUBE FOR USER-GENERATED GEODATA

2.1 The Open Data Cube

The ODC is a Python-based free and open-source software released under the Apache 2.0 license. This software works as an intermediary layer between satellite data and the users (Open Data Cube, 2021). It is designed to reduce the data pre-processing and processing effort required for the harmonization and integration of different data sources and formats. The ODC provides tools for data exploration, processing, and analysis which obey the Open Geospatial Consortium (OGC) standards (Killough, 2018).

The currently operative ODC implementations provide end-users with tools for handling a variety of Earth observation data and derived products. An example of this is the Swiss Data Cube (Swiss Data Cube, 2021), a web ODC instance dedicated to the exploration and exploitation of satellite Sentinel-2 and Landsat-8 imagery acquired over Switzerland. Other relevant ODC local instances are the Vietnam Open Data Cube (Vietnam Open Data Cube, 2021), Digital Earth Africa (Digital Earth Africa, 2021), and the Digital Earth Australia (Digital Earth Australia, 2021) which have been developed to favour the access and exploitation of Earth observation data at a national or continental level.

Actually, the existing ODC deployments support only Analysis-Ready Data (ARD), which are available from a few Earth observation missions (Giuliani et al., 2019). As an example, the Copernicus Programme provides both ARD (e.g. Sentinel-2) and non-ARD (e.g. Sentinel-5P) and therefore only the former can be integrated directly into the ODC through the available ODC data ingestion pipelines. Different data sources and formats cannot be ingested into the ODC without extensive preprocessing and formatting operations. In this work, the ODC is employed to ingest and manipulate grid data retrieved from non-satellite data sources, such as the Facebook users' population maps which are described in the next section. The development and application of a data ingestion pipeline to extend the capabilities of the ODC also to this unconventional data source are described in section 2.3.

2.2 Facebook users population maps

The user-generated geodata used in this study is obtained from the Facebook Data for Good initiative (Facebook, 2021). The initiative provides policymakers and researchers with space-time resolved geolocated information on Facebook usage in areas affected by crisis events. As an example, this data is useful for depicting how the population displace in response to the disaster. The dataset called Facebook Population Maps was considered in this study. These maps provide counts of Facebook users, having location services enabled on their mobile devices, that stay within a specific location in a defined time interval. Data locations refer to pixels of geographic grids derived from the Bing© tile system (see Figure 1b). The count of users per pixel is provided with a temporal resolution of 8 hours. In order to ensure privacy protection and to prevent user tracking, spatial and temporal aggregation of individual records, and removal of tiles with small count values, are performed by the data provider.

The dataset is distributed in CSV format in the WGS84 geographic reference system and can be downloaded by registered users from the Facebook Data for Good partner portal (https://partners.facebook.com/data_for_good). The spatial grid resolution for the Insubria region was approximately 850 m (see Figure 1b). Data of the period May 6th 2020 - December 31st 2021 were considered in this study (see Figure 2). One CSV file for the whole region every 8 hours is available. Accordingly, three values per day - in the time slots 12 am-8 am, 8 am-4 pm, and 4 pm-12 am Greenwhich Mean Time (GMT) - are provided at each pixel location.

2.3 Facebook data ingestion into the Open Data Cube

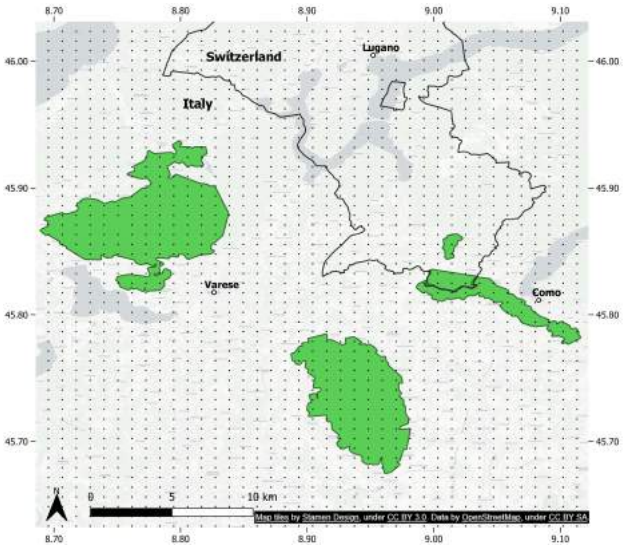
As anticipated in section 2.1, the ODC was used to here ingest and register raw CSV Facebook Population Maps data files into a PostgreSQL database, which is used by the ODC system for data storage. This procedure is required to index data records and allow the system for extracting data upon user queries in an array-like format.

The first operation for the ingestion was the registration of a new data *Product* in the ODC. The Product is a functional object of the data cube that serves as a grouping layer. This object does not contain any data itself, but it is used as a reference to define connections and assign metadata to different datasets. The Product is the key object on which query operations are based.

The second operation consisted of the development of a custom Python processing pipeline used to pre-process each CSV file. This pipeline provides two outputs. The first one is a NetCDF file containing the data which can be handled with the Python XArray library. The second is a metadata file in YAML format that describes attributes of the data such as Coordinate Reference System (CRS), resolution, format, and the type of content. Once both of these files are created, the pipeline registers them on the ODC PostgreSQL database. During the registration operation, the data is assigned to a Product. It is important to note that this operation is repeated for each of the thousands of CSV files used in this work, hence the importance of creating an automated pipeline.

Access to and space-time queries on data were then enabled by the ODC Application Programming Interface (API). Data querying in the ODC can be performed on any of the space-time dimensions contained in the data by specifying its related Product name. The results of these queries are Python XArray datasets that can be directly used e.g. for exploratory statistical analysis of geolocated time series of observations, such as the case of the present work (see e.g. figures 2 and 3).

(a)



(b)

Figure 1. The natural parks of the Insubria Region (a) and the Bing tile grid centres position for the study region with the main urban centres (b). Basemaps: (a) Bing Satellite © 2022 Microsoft, (b) Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY SA.

3. CASE STUDY: VISITORS TURNOUT ANALYSIS IN THE INSUBRIA NATURAL PARKS

3.1 Data processing and analysis

Experiments were run by aggregating Facebook users' counts time series both in space and time. Spatial aggregation consisted of grouping users' counts by parks, in order to assess users' presence across the different parks in the study region. Time aggregation was performed to analyse the dynamics of users' presence for different days of the week and to quantify differences in terms of mean users' counts between weekdays and weekends in the different park areas. Patterns across space and time were considered to outline local differences between parks and surrounding areas.

The exploration of Facebook users' count differences between weekends and weekdays was accomplished by computing the relative changes between the mean users' populations for the different parks in the study region. These relative changes were considered as an indicator of the use of natural parks for leisure or tourism activities that are expected to be more intense during weekends for the park areas compared to the surrounding urban areas.

A synthetic indicator describing local peaks of users' presence was extracted from the users' count times series and mapped to visually identify the most visited locations within the parks (Vavassori et al., 2022). The indicator was expressed as the ratio between the 90th percentile of the day-time (8 am-4 pm and 4 pm-12 am GMT) Facebook users' counts and a baseline value for each location. The baseline was computed as the mean of Facebook users 'counts during night-time (12 am-8 am GMT) which was used as an estimate of the fraction of the resident Facebook users' population. The 90th percentile was arbitrarily chosen as the peak value instead of the maximum value to portray more frequent peaks of users' presence and not only possible extraordinary peaks produced by sports events, festivals, etc.

Data processing and analysis were performed by coupling Python data analysis libraries, such as Pandas (https://pandas.pydata.org), Geopandas (https://geopandas.org), and Seaborn (https://seaborn.pydata.org) with the ODC API. Custom functions were developed for data aggregation, statistics computation, and results plotting. Results were summarized with graphs and maps as presented in the following section.

3.2 Results

Results allowed exploring similarities and differences between parks in terms of visitors' turnout space-time patterns. The recreational value of natural parks was disclosed by analysing the daily mean users' counts time series as well as the increases/decreases of users' presence during weekends compared to weekdays. Furthermore, locations affected by the highest peaks of users' presence were detected by mapping the local indicator (i.e the ratio between the 90th percentile and baseline users' population values at each location) described in the previous section.

Facebook data representativeness for the study area was preliminary computed as the ratio between the mean night-time Facebook users' counts and actual resident population retrieved from the WorldPop dataset (WorldPop, 2021). Average representativeness of about 5% was identified (Oxoli and Brovelli, 2021). The numerical disclosure of results and their use in management operations shall always consider this limitation.

The first comparison between mean Facebook users' counts for each park and the surrounding no-park area was carried out through the spatial aggregation of single-locations time series. Figure 2 represents the time series of the daily mean users' population for each park and surrounding no-park area. The magnitude of mean users' counts is marginally proportional to the spatial extent of the corresponding natural park (see Figure 1b). The effect of the summer holiday period (generally August) is visible as a drop in the time series during both years 2020 and 2021. The summer drops in users' counts are evident in each natural park area but particularly evident in the no-park area, which also includes large urban centres.

Furthermore, all the time series exhibit a weekly seasonality. Peaks of Facebook users' population occur during weekdays within the Swiss natural parks (Parco del Penz and Parco Gole della Breggia) as well as in the no-park area. On the opposite,

the Italian natural parks (Parco Campo dei Fiori and Parco Pineta) show peaks in users' presence during weekends. Way lower weekly excursion of mean users' counts is exhibited by Parco Spina Verde, where population dynamics may be affected by the nearby Como urban centre during weekdays and by visitors turnout during weekends. This effect can be also due to the spatial resolution of the data that, particularly in the case of Parco Spina Verde, results in overlapping Bing tiles between the park and urban centres area (see Figure 1b) thus biasing the patterns which cannot be investigated within the park boundaries only. The same effect is likely to show up in the case of Swiss parks due to their relatively small extension as well as proximity to urban centres.

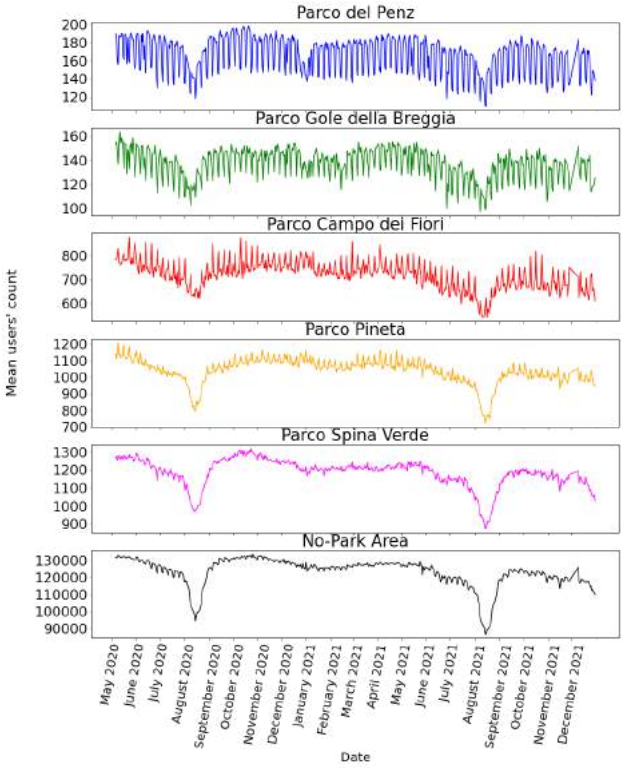


Figure 2. Time series of daily mean Facebook users' counts by park for the study period May 6th May 2020 - December 31st 2021.

To investigate the weekly variations of the time series and differences in mean Facebook users' population between weekdays and weekends, the time series were aggregated by day of the week as shown in Figure 3. Such an aggregation allowed disclosing influences of short-stay tourism (i.e. day trips or weekends) which are expected to happen mostly on not-working days.

Population patterns across the different areas follow those observed and discussed in the previous comparison. All park areas, as well as the surrounding areas, show mean Facebook users' counts with no significant variations during weekdays. Mean users' population tends to increase during weekdays across Parco Campo dei Fiori and Parco Pineta by pointing out visitors turnout increases for leisure activities in these major parks for the Italian side of the Insubria Region.

On the opposite, patterns for the Swiss natural parks (Parco del Penz and Parco Gole della Breggia) show a decrease in users' presence during weekends, suggesting these areas as less frequented destinations for short-stay leisure activities than the Italian side. The same consideration applies to areas outside the parks. Indeed, users' population dynamics in the parks neighbouring areas are more complex to disclose, as the region outside the parks' boundaries includes a mix of large urban centres, forests, rural areas and other famous tourism destinations, such as sub-alpine lakes.

Lastly, the mean users' counts within Parco Spina Verde exhibit small variations over the different days of the week. Unlike the other areas, this natural park stretches close to Como urban centre. Therefore, users' population dynamics are expected to be largely affected by urban users' population.

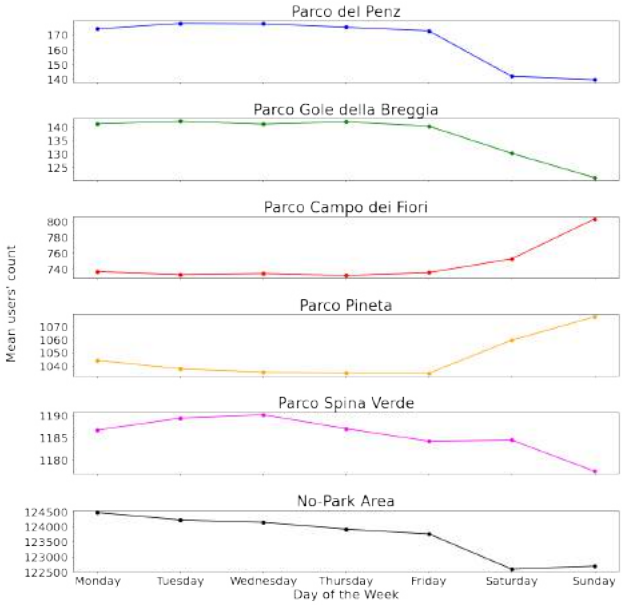


Figure 3. Variation of Facebook mean users' population count per day of the week.

The computation of the relative changes in mean users' population between weekends and weekdays provided additional pieces of evidence on the similarity of the population patterns between Parco Campo dei Fiori and Parco Pineta while remarking their differences from the other park areas. Table 1 shows the relative changes between weekdays and weekends. The relative change was computed as the difference of the mean Facebook users' population between weekends and weekdays, normalised to the weekdays mean users' counts.

The most significant increase is experienced by Parco Campo dei Fiori area (+6%), followed by Parco Pineta (+3%), providing insight into their appeal for recreational and leisure weekends activities. On the opposite, a relevant decrease during weekends compared to weekdays is exhibited in Parco del Penz (-20%) and Parco Gole della Breggia (-11%) areas, suggesting the two Swiss parks are less frequented tourist destinations if compared to the other natural parks. The relative change is close to zero for Parco Spina Verde natural area, highlighting no significant differences in mean users' counts between weekends and weekdays. A small decrease (-1%) of relative change is exhibited by the no-park area, in which both yearly and weekly users' population dynamics are driven mainly by the large urban centres of the study region.

Furthering the data exploration, peaks of users' presence compared to the fraction of the resident population were assessed by computing the indicator described in Section 3. The indicator was computed and mapped at each pixel location (Bing

	Relative changes between weekends and weekdays [%]
Parco del Penz	-19.71
Parco Gole della Breggia	-11.06
Parco Campo dei Fiori	5.98
Parco Pineta	3.01
Parco Spina Verde	-0.56
No-Park Area	-1.18

Table 1. Relative changes [%] of average users' population between weekends and weekdays.

tile centres) within the parks' boundaries. Figure 4 represents the spatial distribution of the ratio between the 90th percentile of mean Facebook users' day-time counts and the baseline value. This result allowed the detection locations affected by the highest increases in users' presence with respect to the resident population.

The indicator map suggested higher increases in users' presence near the borders of the natural parks. The result is reasonable as these areas are near to and accessible from the surrounding urban centres. However, the scarcity of data in the inner parts of Parco Campo dei Fiori and Parco Pineta did not allow for a consistent assessment of the indicator across the whole natural parks areas. The lack of data may be due either to the limited Internet connection or to an actual lower visitors flux related to more limited accessibility or attractiveness of these areas than the outer areas.

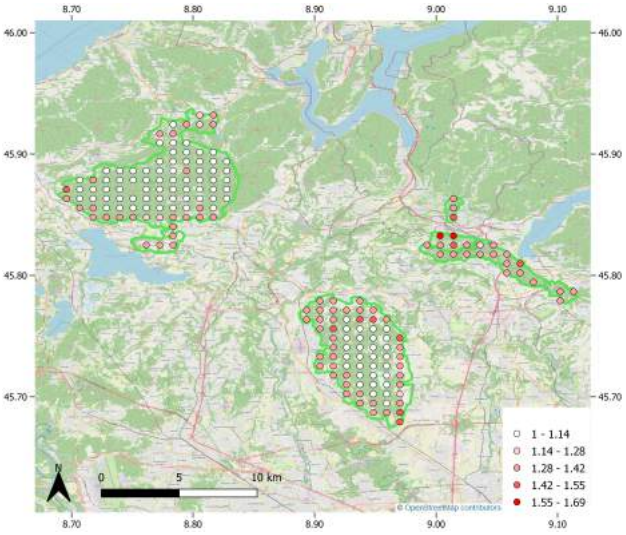


Figure 4. Map of the ratios [%] between the 90th percentile and the baseline line values (night-time records) of the time series of Facebook users' counts. The ratio is computed at each location (Bing tile centre) within parks and used as a metric to infer the most visited locations (Basemap:© OpenStreetMap contributors)

4. CONCLUSIONS

In this paper, an investigation of visitors turnout in the natural parks of Insubria Region, between the Canton Ticino (Southern Switzerland) and the Lombardy Region (Northern Italy), was presented. Visitors turnout space-time patterns and trends were assessed by leveraging user-generated geodata provided by the Facebook Data for Good initiative for the study period May 6th 2020 - December 31st 2021. As one of the novelties presented in this work, the application of a cutting-edge data management software tool (the ODC) was presented for the manipulation and integration of a large volume of geospatial and temporal data.

Specifically, the Facebook Population Maps dataset was used in this study. The dataset consisted of geolocated time series of Facebook users' counts, allowing for the assessment of space-time users' population patterns between parks and surrounding areas.

Exploratory analyses were carried out to detect users' patterns and trends for each day of the week, pointing out differences between weekdays and weekends. Numerical outcomes were represented with graphs and tables, providing a tool to infer visitors turnout in the study area and assess the tourism attractiveness of natural parks compared to the surrounding urban and rural areas. A space-resolved local indicator was computed and mapped to point out most visited destinations within the natural parks.

Results outlined similarities between Italian parks (especially Parco Campo dei Fiori and Parco Pineta) with increases in mean users' presence during weekends, suggesting patterns being significantly affected by short-stay tourism fluxes. An opposite pattern was detected within the Swiss parks (Penz and Gole della Breggia), as well as in the areas outside the parks. In this case, the decrease in users' presence during weekends suggested these areas be less popular destinations. A peculiar users population pattern, with a significant drop in users' presence during summer and limited weekly variations, was detected for the Parco Spina Verde area. In this case, users' dynamics is suspected to be significantly affected by the proximity to the Como urban centre.

The map of the ratio between the 90th percentile of mean users' counts to the baseline users' count values allowed for the detection of locations affected by the highest increases of visitors presence within the natural parks. The most significant peaks of users' presence were detected close to the borders of the natural parks. However, this result may be partially biased by the scarcity of data or by the reduced accessibility of the inner parts of some of the natural parks.

The ODC system was here tested to manage non-conventional, large volume datasets, namely the Facebook Population Maps. Despite being originally meant to handle Earth Observation satellite images, the ODC proved to be an effective tool for storing, processing, and analysing also the considered user-generated geodata. The ODC software provided a cloud-computing framework for handling large volumes of data with a single end-point. Furthermore, the exclusive use of free and open-source software provides the analysis with the potential to be replicated and improved.

As a final remark, a critical issue when dealing with user-generated geodata is connected to data representativeness. Facebook data representativeness is questionable owing to the size of the users' sample which is a limited portion of the actual population. The representativeness expressed as the ratio between average Facebook users' counts and the actual resident population (Oxoli and Brovelli, 2021) resulted in about 5% . However, the user-generated geodata provided by the Facebook Population Maps proved to be an unequalled source of information in terms of availability, spatial resolution, and temporal frequency. Comparable information would not be possible to achieve with traditional surveys or interviews.

The obtained results represent valuable information for park managers and local stakeholders to best understand seasonal and weekly visitors' fluxes variations and to take evidence-based local management actions. Future developments of the work will focus on the definition of new analyses to extract additional details useful for a better explanation of the visitors' turnout by considering, comparing, and integrating ancillary data on local tourism fluxes. It is worth noticing that the study region was affected by COVID-19 mobility restrictions during the study period. Therefore the analysis is expected to be replicated on an extended time period to investigate also variations due to this extraordinary situation.

REFERENCES

Di Minin, E., Tenkanen, H., Toivonen, T., 2015. Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, 3, 63.

Digital Earth Africa, 2021. Digital Earth Africa. (https://www.digitalearthafrica.org. Accessed 29.05.2021.

Digital Earth Australia, 2021. The Digital Earth Australia. https://www.ga.gov.au/dea. Accessed 29.05.2021.

Facebook, 2021. The Facebook Data for Good initiative. https://dataforgood.fb.com. Accessed 29.12.2021.

Giuliani, G., Camara, G., Killough, B., Minchin, S., 2019. Earth observation open science: enhancing reproducible science using data cubes. *Data*, 4(4), 147.

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., Di Minin, E., 2018. Social Media Data Can Be Used to Understand Tourists' Preferences for Nature-Based Experiences in Protected Areas. *Conserv. Lett.*, 11(1), e12343.

Killough, B., 2018. Overview of the Open Data Cube initiative. *Proceedings of the 2018 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 8629–8632.

Markevych, I., Schoierer, J., Hartig, T., Chudnovsky, A., Hystad, P., Dzhambov, A. M., De Vries, S., Triguero-Mas, M., Brauer, M., Nieuwenhuijsen, M. J. et al., 2017. Exploring Pathways Linking Greenspace to Health: Theoretical and Methodological Guidance. *Environ. Res.*, 158, 301–317.

Open Data Cube, 2021. The Open Data Cube. https://www.opendatacube.org. Accessed 29.05.2021.

Oxoli, D., Brovelli, M., 2021. Citizen-Generated Geodata for Natural Parks Use Analysis: Insights from Facebook in the Insubria Region. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, 43, 195–199.

Oxoli, D., Cannata, M., Terza, V., Brovelli, M., 2019. Natural Heritage Management and Promotion Through Free and Open Source Software: a Preliminary System Design for the Insubriparks Project. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, 4214, 179–183.

Prell, C., Hubacek, K., Reed, M., 2016. Stakeholder analysis and social network analysis in natural resource management. *Handbook of applied system science*, Routledge, 367–383.

Swiss Data Cube, 2021. The Swiss Data Cube. https://www.swissdatacube.org. Accessed 29.05.2021.

Vavassori, A., Oxoli, D., Brovelli, M. A., 2022. Population space-time patterns analysis and anthropic pressure assessment of the Insubric lakes using user-generated geodata. *ISPRS International Journal of Geo-Information*, 11(3), 206.

Vietnam Open Data Cube, 2021. The Vietnam Open Data Cube. https://datacube.vn. Accessed 29.05.2021.

Wood, M., 2002. *Ecotourism: Principles, practices and policies for sustainability*. UNEp.

Wood, S., Guerry, A., Silver, J., Lacayo, M., 2013. Using social media to quantify nature-based tourism and recreation. *Scientific Reports*, 3, 2976.

WorldPop, 2021. Population counts. https://www.worldpop.org/geodata/listing?id=78. Accessed 30.07.2021.