Masters Program in Geospatial Technologies

IDENTIFY POPULAR HOTSPOTS THROUGH THE ANALYSIS OF MOVEMENT PATTERNS FROM SOCIAL NETWORKS IN RURAL AREAS

Case study of the Borbera Valley in North Italy

Stefano Cunietti

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

NOVA IMS Information Management School
Universitat Jaume I
Westfälische Wilhelms-Universität Münster

Identify popular hotspots through the analysis of movement patterns from social networks in rural areas

Case study of the Borbera Valley in North Italy

Stefano Cunietti

Student at the Universitat Jaume I, 12071 Castellón, Spain
al422314@uji.es
Student ID: 422314

Identify popular hotspots through the analysis of movement patterns from social networks in rural areas

Case study of the Borbera Valley in North Italy

Dissertation supervised by:
Joaquín Torres Sospedra, Ph.D. Institute of New Imaging Technologies (INIT)

Universitat Jaume I, Castellón, Spain

Dissertation co-supervised by:
Sergi Trilles Oliver, Ph.D. Postdoc at Universitat Jaume I, Geospatial Technologies Research Group, Institute of New Imaging Technologies (INIT)

Universitat Jaume I, Castellón, Spain

Dissertation co-supervised by:

Flávio L. Pinheiro, Assistant Professor at NOVA IMS (Information Management School)

Universidade Nova de Lisboa, Portugal

February 2022

DECLARATION OF ORIGINALITY

I declare that this work has been produced entirely by me referring to works previously carried out on same topics connected to this thesis and that these works have been cited when reference has been needed.

All the data used were made public and visible by users, if you find something does not correct, please contact me at the email: m20200031@novaims.unl.pt or niettiste@hotmail.com I will delete what you do not find should be showed anymore.

This work has not been previously evaluated and/or submitted to the Universitat Jaume I (UJI), Castellón, Dept. Lenguajes y Sistemas Informaticos (LSI), Castellón, Spain.

Castellón, Spain, February 2022
Stefano Cunietti

ACKNOWLEDGMENTS

First, I want to say how important it is for me to have reached this milestone for my life and the ransom I had from it. I have never stressed this fact, but I come from a small village in the countryside where at the end of the 90s, many things were not taken for granted, where knowing at least Italian correctly was already an important step (most of the people spoke dialect). Many teachers did not believe in me and indeed many times, they discouraged me from making choices because they did not judge me as capable of achieving good academic goals. The school results since elementary school in my small countryside village school have never been satisfactory, no one has ever tried to understand what my discomforts were, taught an efficient and effective study method.

Fortunately, I continued in my decision even when the family did not support or did not understand my idea, especially later to continue studying in this master. I learned for myself and what was the best method of study or at least the one that allowed me to have results, often alone and disheartened. Many of these steps were taken during the first year of university, where I also had to run to reach the level of the others.

I am very grateful to my old university, the Polytechnic of Turin, which gave me the opportunity to make important experiences, including my first exchange abroad in Freiburg in Germany, a city that I always carry in my heart, where I had a deeper knowledge of geospatial technologies and gave me the energy and the desire to get here.

Thanks to this experience I learned English and not only that, I had the strength to have further experiences abroad some fruitful others not, but in any case they helped me to grow.

I want to especially thank Professor Marco Painho. Despite this complicated period that we are experiencing, it was nice to be able to attend university classes in person in the first semester. I will always remember the NOVA campus as a refuge in this time of great uncertainty and I thank the professor for allowing us to find such a situation, a privilege when most of the other classes were taking lessons from home. This also allowed us to establish relationships, exchange ideas, visit places or simply spend moments of relaxation with classmates.

I really thank Professor Joaquín Torres Sospedra who not even knowing me has agreed to follow me in this project together with Ph.D. Sergi Trilles Oliver from the Universitat Jaume I and Flavio Pinheiro Invited Assistant Professor at NOVA IMS Information Management School all with great patience.

Identify popular hotspots through the analysis of movement patterns from social networks in rural areas

Case study of the Borbera Valley in North Italy

Abstract

Social networks are now an increasingly used tool, but analysis possibilities have not yet been fully exploited. In particular, the extraction of information from users' profiles and their processing could give different information. In this work we will focus on the possibilities of using this information to analyse the patterns of rural spaces. The work will be carried out through a review of the available bibliography on the topic, the construction of an application, and the subsequent analysis of the data extracted through the application. Based on the findings, suggestions are made about the intensity of people within an area or the changes that have occurred in social activities.

Keywods: social medias, social medias scraping, Instagram, network analysis, Sklearn, GEPHI

Summary

List of the figures

List of the tables

1 Introduction

In the digital age, we live in everything is interconnected. From personal devices owned by individuals or companies to public that generate, accumulate, and share data. These elements create the basis for what is often referred to as the Internet of Things (IoT). Regardless of our views on data privacy and data ownership, the fact is that our data is being constantly used, consciously or not, and analysed with different goals and purposes to bring added value to third parties. Such analytical processes are only possible due to new methods and operations that fall within the realm of the so called "Big Data Analytics" (Babar et al., 2017). Through this combination of IoT and Big Data Analytics, we could gain such a large amount of useful information to derive new knowledge and insights on phenomena relevant to many sectors of our society. Here, we focus particularly on the branch of IoT and Big Data Analytics applications to urban planning (Rathore et al., 2016).

Of course, there are not only benefits (Lee et al., 2015), those who have access to this information may have an advantage in the market or can simply sell it not always for positive purposes (Rodriguez et al., 2012).

1.1 Motivation, and background

New technologies are completely changing how we interface with the world and people. Among all the tools made available by the digital revolution, there are social media. They are still used with great diffidence in academic research due to the specific targeting of users. However, the situation changes due to the increase in users, making the investigation samples less polarized, and developing new analysis techniques (Camacho et al., 2020). The statement is testified by the fact that searching for academic articles in support of this work, it was difficult to find projects with a high number of citations and used in solid projects, but above all that have been peer to peer reviewed.

The increase in users means sociability is more and more happening through social medias, whereas it was once essential to have face-to-face relationships (Barkhuus et al., 2010). In fact, they are becoming always more a popular phenomenon reaching the entire global population. It is increasingly common to share moments of life through these channels, so much that part of the physical life has been transferred within these platforms. As a result, digital has become part of our existence, through social media, we not only share what we do, but we use them to know or publish events, read news or to interact with people, for simple friendship or even for business (Hudson et al., 2015).

Therefore, we can have knowledge of the world, of events and everything happens, places to visit, without having more direct relationships with people. In some cases, it may be essential to have these tools to be able to start a face-to-face interaction, to be able to show oneself to others above all strangers. An example could be a student who is in the first year of university, and she/he wants to participate in events organized by others (Ibáñez-Cubillas et al., 2017).

This leads not only to positive impacts but negatives too. People can narrow their knowledge field only to what is of interest, completely isolating themselves from the rest (Gosal et al., 2019). Of course, there are many other factors that can influence people such as culture, level of education, origin, social class, and gender (Candia et al., 2019).

Often data analysis from social media does not bring any popular advantage, but the companies use them even by whom own those platforms to do business. This is because they are using them themselves or reselling data got to other companies to understand the direction of the market or even by targeting through advertising (Kim et al., 2012).

Referring to the urban analysis, there is often a lack of data on which to base the work, and the extraction of data from social networks could be a solution. As we have said that these platforms have a massive impact on people's sociability, this bring on a different way of living spaces and for this reason, it is important to understand and study them (Martí et al., 2019). Just think of the influence it can have in directing people to a specific place, thus triggering a significant tourist influx (Oliveira et al., 2015).

Social media can be a solution even if it depends on case by case and on the specific motivation; the challenge is to be able to develop a methodology to ensure that what we want to know is correctly reported by the data collected from them (Boy et al., 2017). In any case, we must always think on the specific characteristic of each platform because even if it changes slightly, this could be particularly useful in a certain field of study. By extracting this data, it could be an important and huge source of study to reveal how people use the spaces and how the city's economic and social dynamics are shaped. An example could be understood how different economical vocations of the city are divided (work, fitness, free time, nightlife, etc.), the influx of people to a place, the environmental quality, language of the people who frequent a place and many other examples as found in the literature which we will discuss shortly.

It is precisely starting from this consideration that this work takes inspiration, verifying whether it is possible to get acquainted from data obtained through online social media platforms, both from locals who live in a region but also from non-locals who might visit the region for many reasons (e.g., visiting friends, tourism, work, etc.).

1.2 Aim, and research questions

Urban planning includes various disciplines, and their interface has to potential to suggest concrete solutions to solve a wide range of socially relevant problems (e.g., social segregation). Each of them is a decision based on data related to this specific area; they are geo-referencing the specific place. Together with geospatial technologies, among all the subjects that make up urban planning, sociology is the one it plays in the author's opinion. Combining this last statement and when said of the previous paragraph, social media could always have greater importance for understanding their dynamics. For this reason, the main interest of this work concerns the possibility of using these unconventional data sources in a consistent way to understand movement patterns in peripheral urbanized areas.

Making particular reference to geospatial technologies, we wanted to try to develop what has been learned in the branch of statistical analysis and programming to obtain a complete process from code to analysis and layouts. To achieve this, the work will start from writing a code to create a program that can obtain data from public Instagram accounts. The goal was to use as few as possible tools outside those offered by the python world, including the return of graphs and maps. Instagram no longer offering its own API has dramatically reduced the possibilities of getting to know through its platform even in the university environment, so that one of the

objectives is also to be able to carry out a job not only in a rural area, but also with reduced possibilities and prolonged times in obtaining data.

The decision to extract information from Instagram derives from the fact that, from an initial search of academic articles, several authors indicated it as the platform that offers the most relevant insights in urban studies (Martí et al., 2019).

In addition to demonstrating what was previously mentioned, it has been verified a lack of consistent works in the literature carried out in peripheral areas, but analysis focused on the urban environment. So, we want to verify the possibility of getting information also for rural areas, where there is no concentration of people as in cities. Therefore, the number of users also decreases drastically. Another impacting factor for the analysis is that usually, in peripheral areas, the number of young people is lower. However, the segment of the population makes the most use of the tools.

In the context of this thesis, it is a question of making an investigation involving knowledge of the location of the place where the post is published. Therefore, these elements can characterize the place, but this also helps to understand what the main activities are, by which target people it is frequented such as the language spoken in the description can be an indicator in tourism of the origin the influx of people.

Finally, we want to apply techniques learned from other works carried out, such as the clustering of users and the classification of the activities carried out by people in the posts.

1.3 Literature review

The difficulties are nowadays many, one of them is the extraction of data, few social networks with certain characteristics make their data available at least for university research, when they do it usually it is platforms that are not popular and/or addressed to a specific target of people (Abdulrahman et al., 2013). Another important question is to verify the validity of the dynamics identified during the analyses. In fact, the social networks even if they are becoming more and more a popular phenomenon involving all age groups and people with different interests, there is no security they always have social balance (Martí et al., 2019).

The literature review was structured as follows:

1. understand how social media can influence people, especially in the knowledge of new places.
2. which social media to use and what potential for understanding the phenomena it could give back, ease of obtaining data and quantity.
3. scraping techniques already applied as an example to use.
4. text classification techniques.
5. Past works analyses already carried out and applied in the academic and public urban planning fields.

1.3.1 Understanding environmental, social, and urban values through social medias

There are studies to understand the effects and influence of social media in various sectors of people's lives: politics, shopping, holiday destinations, choice of restaurant, visiting places or participating in events of cultural interest (concert, museum, etc.). Several academic papers have

been found to measure these phenomena. Among those, an attempt was made to identify those who, within their analysis process, had included a data scraping phase and have focused on socio-economic aspects.

What was immediately understood by the review phase is that mainly social media such as Instagram, Facebook, Twitter, can be useful tools for analysing people's habits. However, at the same time, they are also what creates them. Therefore, some studies on understanding people's preferences during their free time were interesting.

Just think of the impact those social media platforms can have in activities such as tourism. Those are also increasingly assuming a central role in Western society, becoming one of the main economic sectors. Social media often play an important role in the choice of destination; people consult or view other profiles or promotion pages (Latorre-Martínez et al., 2014). It is even possible to study the language used in describing the action expressed within the publication to understand what the user was doing. The words used referring to the same action may also indicate, for example, how they can be linked in different languages (Ronen et al., 2014).

Some papers tried to gather information on the gastronomic culture of the city of Macau and the appreciation of the dish or not of the people who had posted (Yu et al., 2019). Some simply dedicated themselves to identifying which locations were the most popular for a specific activity like restaurants (Zhai et al., 2015). Even those have started practising a real branding of a place with these powerful means, which was partly successful (Oliveira et al., 2015). Still others to measure the extent of product sales (Itani et al., 2017).

Some works focused on capturing the value that users want to express through photos. One of these gives a general view on the methodologies to be used and why to choose a platform according to the final research objective. It also recommends using similar but different platforms to validate the data (Martí et al., 2019). By verifying the most frequented places around a river, indicated by a high number of publications, a project tried to verify where it was more suitable to build a dam because it can radically change a landscape and its context (Chen et al., 2018). Using the same principle, another project tried to verify which policies would be useful or not to undertake to implement green areas in the city of Copenhagen. For this reason, citizens were asked to publish a photo with a specific hashtag so that the images could be retrieved and analysed (Guerrero et al., 2016).

There are studies to understand the social structure mainly in cities moving in the social sphere. The first through all the information made available by Instagram posts, they tried to comprehend how the various areas of the city were experienced by people (Boy et al., 2017). The location was useful to see the intensity with which people frequent the places, the hashtags, and the description to understand the type of activity and even give greater or lesser importance to certain information, comments, and likes. Another academic paper used a similar method to the previous one to measure social diversity within the city of London, verifying how people accessed different urban spaces (Hristova et al., 2016). Another project has further developed the themes already stated, clustering people into groups by interaction and intensity. It islso localized for the place to measure the level and nature of segregation between different social groups (Boy et al., 2016). There are also studies on understanding large mass events such as the "Arab Spring" whose call was spread through the social media (AlSayyad et al., 2015).

1.3.2 Which social media to scrape, limits and possibilities

Among the main reasons for choosing to use Instagram is the possibility of having a high number of posts located in a specific place (Domínguez et al., 2017), differently from other social media such as Facebook or Twitter. Another advantage that made this platform preferable over others with similar characteristics, such as Flick, is that it has now become popular and used by people with different interests, while the latter has a very targeted audience (Martí et al., 2019).

Anyway, among all the technologies currently available for sharing content, one of the most popular is Instagram; launched in 2010 it has now reached an estimated number of users of nearly one billion.

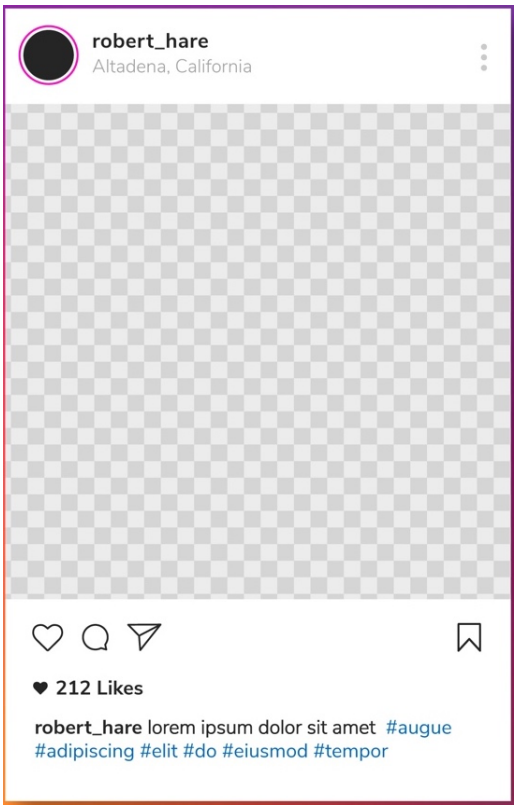

Figure 1. Designed by Freepik http://www.freepik.com

On Instagram, commonly users publish a “post” (Figure 1) which is the term used to indicate the set of contents that can be inserted following a pattern defined by the platform, even if it is not necessary to insert all of them, but some are mandatory and are:

- **Picture** (mandatory), it is the main content that all information is about.
- **User nickname** (automatically inserted from the platform).
- **Location** (not mandatory), it can be entered by the user based on the position in which it is publishing, or it can be searched, it is usually referred to the place of the photo.

- **Date** (automatically inserted from the platform), referring to the time the post is published.
- **Description** (not mandatory), is the caption of the picture in which you can insert a text of the desired length and in turn can have:
    - **Hashtags**,
    - **Tags of the people present in the picture**.

Once the post is published, other users can interact with it by liking or commenting. The platform then shows them, and you can also see the popularity with the number of users who have like or commented.

The schematization in this way for publication is conducive to use in academic research compared to other freer platforms previously mentioned. In fact, unlike what happens for Facebook or Twitter where the user has greater freedom to customize the post to be published, the main feature of Instagram is that the only content always present and that can be shared through a post is the image contained within it (Martí et al., 2019). In some cases, this content is useful to understand what value is given by those who take the photo relatives to a specific place, both at a social level and at a landscape level. However, it is impossible to have information about the user in a standardized way (age, gender, origin, level of education, etc.) even if you want, as it happens for Facebook.

One of the things immediately revealed using platforms for searching academic papers globally, but also from other platforms, was the fact of having some relatively prosperous years in the field of scraping and data analysis and then suddenly stop. Starting from June 29, 2020, the information with which the academical papers obtained the data before that time are no longer entirely valid for the extraction part (*Instagram - Wikipedia*, n.d.). Facebook acquiring the platform has standardized the policies, and now it is possible to get information only through two official apps: Instagram Basic Display API and Instagram Graph API (*Instagram Platform*, n.d.). However, these applications only allow to download data about the owner's account and other accounts upon request for authorization. This has created a strong gap in the possibility of replicating what has already been done previously, for this reason new alternative means to acquire data must be found. This leads to the impossibility of replicating analysis on such a large dataset due to the new more stringent policies, limiting the possibility of extracting data both through its official channels and using third-party apps by blocking the account in case it detects misuse.

A solution found was to develop an application and only download information from public posts and in limited numbers. Other developers built some APIs for research purposes, but after several attempts, they did not extract the amount of information available most of the time (for example, the real total of the posts was not displayed, this is because it was necessary to scroll to the end of all the posts related to a place or a hashtag).

Despite the solutions found, the limitations do not allow to use Instagram and the information contained in the posts at a higher consistence as done in other works. Furthermore, in data scraping, attention must be paid to how the data is organized regarding the privacy policies desired by the platform (Bello-Orgaz et al., 2016). The Facebook policies do not make possible to download the photos, as consequence could not be applied any machine learning techniques

combined with the textual description for an advanced knowledge of the activity carried out by the user.

Due to the limitations mentioned above, many of the academic papers found during the literature review prefer to use other social networks such as Twitter and Flickr instead of Instagram (Mamei et al., 2010). Twitter adapts to the analysis application in different sectors since it provides an API and it is possible to obtain different information (location, hashtags, description). However, the posts do not always contain images or other informative content, while Flickr gives the possibility to geotag the photo where it is taken.

1.3.3 Scraping techniques

For developing research like this, the investigation of academic articles is important not only to understand theoretical aspects of the subject matter but also to try to acknowledge practical application processes in a general way. This is to verify if it is possible to find real dynamics that occur in urban contexts and human spaces through social media.

From a first reading, there are already several methods for extracting information described on the web or in academic papers. The word to indicate the process for obtaining data on social networks and in general on the web is: “scraping”. Around this word, the phase of literature review for scraping techniques research has been concentrated, together with others such as: “social networks mining”, “social media mining”, “urban analysis social networks”, “urban analysis social medias”, “Instagram scraping”, “machine-learning applied on social media data”, “manage big-data from social medias”.

It has been seen many academical articles describe the process used, but not the code or program that was used or many of the things are no longer valid due to the change in regulations. As already mentioned in the paragraphs before, there are different privacy policies depending on the social network. These do not allow the extraction of personal data by limiting the possibilities of use for research. One of the social networks detected by bibliographies with greater ease of extraction is Twitter, while for Instagram and Facebook there are greater difficulties due to the restrictions imposed by the platforms. So, one of the main objectives for data scraping is to understand how to extract data respecting these limitations. Due to them, it is important to have an analysis objective before the coding part to structure all the work around it. In fact, writing the program and extracting the data took a long time, and if there was an error in saving or processing the data, the whole process had to be done again.

Among the several papers carried out regarding Social Media scraping, some of them are precisely focused on using data for a particular urban analysis. However, very few makes the code or process used available. Of all the articles published on the subject, the most useful met was a thesis titledInsta-Turin” (Aimar, 2020). Anyway, it was not enough. Since the readings are concentrated above all on the descriptive and non-technical part of scraping, it has been necessary to find some specific application examples: GitHub was the main source of information possible to find different codes made by private developers. Three repositories were particularly interesting:

- Instagram Scraper (arc298 / instagram-scraper, 2021),
- Instascrape (chris-greening / instascrape, 2021),
- get-location-instagram (tikseniia / get-location-instagram, 2021).

The result was the writing of a scraper in the Python programming language, divided into several parts which will be explained in the chapters on methodology. The writing was done to simplify the understanding of the code and avoid writing errors, and due to the limited number of requests that can be made to the Instagram server.

1.3.4 Text Classification using Scikit-learn

This paragraph was added after noting most of the posts published and collected through the scraping process were mostly localized and with a description, but also that through the various phases made to have an initial data understanding, cleaning the data the sample to be analysed would be greatly reduced.

Why is it important to state this? Within the description of the posts, even when the geotag was missing, the name of the place was written, or even different or same users itemized the same words to refer to the same specific location. Hence was born the idea of trying to create a model that could predict the location even if the location was missing. Within the Python world, there is a wide choice of packages that can be used for this purpose. After a comparison with others such as Keras, the one chosen was Sklearn, for its ease of use, the accuracy achieved, and the calculation speed.

Sklearn is not a highly advanced library as mentioned in several academic papers that have talked about this tool or even used it for their own works (Szymánski et al., 2019), where they explain its different potentials also weaknesses. One of these lacks was the impossibility of ensembles, which have however been partially overcome in the latest versions. Attempts were made in that case the accuracy was not beneficial, this is due to the lack of descriptions with a precise structure, being the user who, depending on the post decides what to include in the description. It is anyway effective in performing simple actions, especially when it comes to making predictions not on multilevel. This is enough for our case where we want to demonstrate the effectiveness of making analyses on social networks that are effective in making decisions.

There have been several interesting applications using Sklearn. The program processes the words within the column and transforms them into numerical factors that will be used for the probability of that word/phrase.

1.3.5 Network analysis

We have seen in the previous paragraphs how social media are a tool for socializing (Ibáñez-Cubillas et al., 2017); this creates visible connections that can be extrapolated through different techniques and based on the possibilities that each platform offers. Most of them are often already used to undertake product marketing actions or direct people to a certain sector of the market (Wang et al., 2012).

Network analysis can be applied to different studies and investigate complex relationships. In our case, we are talking about Network analysis applied to social media and there are some works already done in this area, usually to understand the social structure (Akhtar, 2014). The idea of carrying out this type of analysis arose from the fact that network analysis was used in various previous works. The thesis on the city of Turin was fundamentally (Aimar, 2020).

Particularly successful was a program called GEPHI, where the creation of two tables to be loaded later, one based on the nodes and another on the interactions that take place between the

nodes called edges, is foreseen (Bastian et al., 2009). Subsequently, it is possible to perform operations on the data entered starting from clustering based on different algorithms and parameters. At the end it is possible to create a layout and modify its characteristics to adapt them to the visualization needs better.

A program called NodeXL makes this type of analysis even easier (Hansen et al., 2011), but it does not seem to have the same functions as GEPHI which are not advanced. Other projects have involved clusteringeoreferenced data with packages that create models directly from the table using programming languages (Cranshaw et al., 2012).

In this case, the aim is to apply this tool to the geo-localized data for each place obtained from the geotags to check if there are correlations between them and which a certain target of people can visit places. Helpful later was the tutorial for creating an advanced layout in GEPHI (Grandjean, 2015).

2 Case study: #valborbera

The availability of open data, at the same time data already processed during the last master's thesis in Urban planning (Cunietti, 2019) and the size of the area, the author's knowledge of the area, made a choice fell on a small valley south of Piedmont, a region of Northern Italy. In particular, the decision was made because the hashtag #valborbera has posts for several years; at the same time, it was well representative of the area and an important quantity of data to be relevant statistically and at the same time comparable with other data mentioned at the beginning of this paragraph.

2.1.1 Geographic location

The Borbera valley is in the extreme south-east of the Piedmont Region (Figure 2), although historically and economically linked to the Liguria region and its capital Genoa. The destiny that led it to be part of the Piedmont Region began a few years before Italian unification with the Rattazzi reform in 1859 (*Legge 23 ottobre 1859 n. 3702 - Wikipedia*, n.d.), detaching this territory and others a little further north and west from the province of Genoa. Despite this, the territory continued to undergo the influence of Genoa, especially during the economic boom that took place after the Second World War, finding itself on the edge of the so-called "industrial triangle", which had at its vertices: Genoa, Milan, and Turin (Felice, 2015).

Figure 2. Location of the study area

Turin was driven by solid industrialization, especially in the mechanical industry where companies such as FIAT drove the sector, Milan was the financial and cultural centre of the country. At the same time, Genoa had large state funding that pushed the economy to develop very quickly and with a consequent very important soil occupation (Amatori, 2013), for the construction of factories and houses for the consequent need for housing with a strong migration

from southern Italy, despite the city being in a small strip of flat land between the sea and the mountains. Therefore, space had to be found for the port and its infrastructures.

When the territory was saturated, the valleys behind the city and the mountains were affected by urban expansion, especially the factories connected here to the port through a dense railway network (Molinari, 1999).

The valley mainly affected by this phenomenon was the one close to the study area, the writing valley. However, it still has the scope of the Val Borbera in its valley floor, where there has been an intense concentration, while the more peripheral villages on the hills and mountains were depopulated, and many have disappeared. Today they are stages on trekking routes.

2.1.2 Morphological and demographic description

The municipalities that are part of the Val Borbera are 10 and are: Vignole e Borghetto Borbera, the main centres of the valley floor and more infrastructured, Stazzano, Cantalupo Ligure, Albera Ligure, Rocchetta Ligure, Roccaforte Ligure, Mongiardino Ligure, Carrega Ligure e Cabella Ligure, the main center of the upper valley (Figure 3).

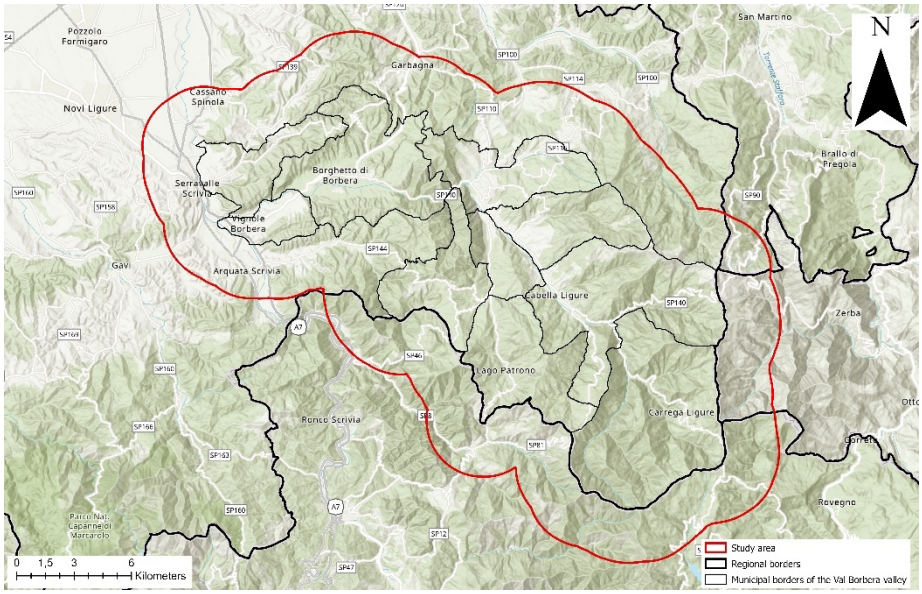

Figure 3. Framing of the territory and infrastructures.

The valley's name derives from the river that flows inside it, called Borbera. Historically, between the valleys of lower Piedmont, they were a place of passage to go to the republic of Genova whose capital was precisely the city of Genova. A city dedicated to commercial maritime traffic and one of the four maritime cities together with Venice, Pisa, and Amalfi. This past is still visible in the monuments and castles of the area, some in good condition and others in the neglect process, along with other military buildings, such as watchtowers, and religious buildings. Together with the geographic morphology of the territory that exceeds even a thousand meters in height, these are destinations for non-mass tourism.

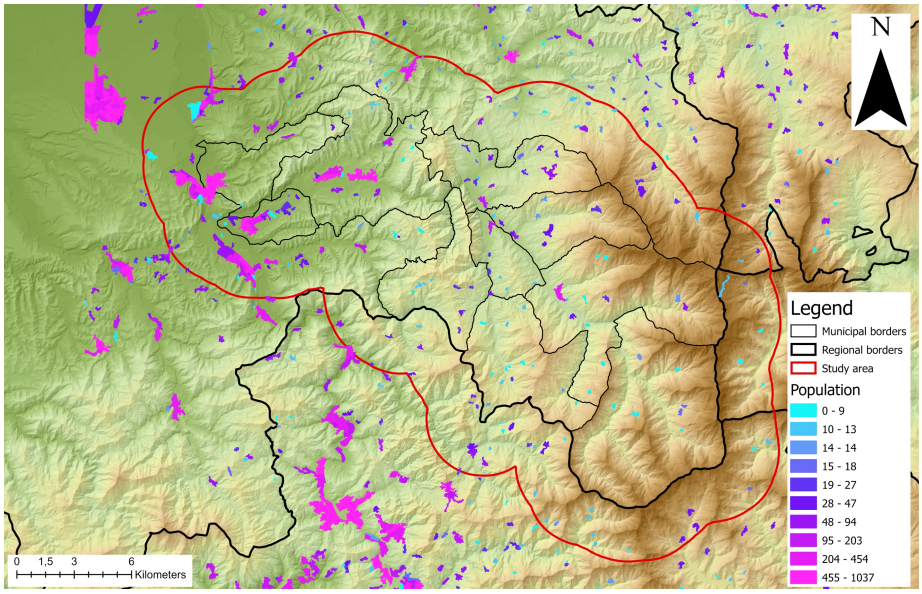

Figure 4. Morphology and concentration of population in the study area

The valley floor is highly infrastructure due to the presence of motorways, state roads and railways that connect Genoa, Milan, and Turin. The situation changes considerably as you go up in altitude towards the upper part of the valley. Except for the municipalities near the access points to these infrastructures, the territory is for the majority still quite isolated due to its geographic conformation. This happens although the area is located on the border between four regions of great economic importance and the strong urbanization of the valley bottom,

The total population of the Val Borbera, the area under consideration, is about 8 thousand inhabitants, which are concentrated mainly in the lower part of the valley. The area under consideration, however, has a higher number of inhabitants about 25 thousand who live in the two towns, Arquata Scrivia and Serravalle Scrivia, immediately outside the valley where in the past, many of the inhabitants of the more peripheral villages moved when they did not emigrate to other European countries or of America.

The territory of the Val Borbera is hilly to the east with some flat parts where the population is concentrated (Vignole e Borghetto Borbera) (Figure 4), increasing the elevation of the reliefs more and more going west where you can also reach peaks of 1300 (Monte Carmo) and 1600 meters (Monte Chiappo). Here is located the only more urbanized centre on the high valley, Cabella Ligure. This area has maintained natural characteristics with several trekking routes to the two peaks and other monuments of the territory, including isolated churches.

3 Methodology

The construction of this thesis goes from collecting data to cleaning and analysing it, producing outputs. The following phases had to deal with the limits on the author's programming knowledge to those imposed by the Instagram platform, which took a long time due to the limited number of requests. Once a certain knowledge on the subject and on the actual possibility of achieving the set objectives was achieved, several other steps were carried out from data collection through coding, inspection, and data analysis.

Due to the need to address these difficulties, the methodology did not follow a linear trend, but having ascertained the opportunity to achieve a certain goal, further analyzes were carried out or were attached. In general, however, to give a more concrete idea of the phases that were used to construct the methodology, it can be summarized as: code writing and correction for scraping, data extraction, data preprocessing, data understanding, producing outputs.

The choice to study an area starting from a hashtag indicated this place has a motivation due to the real possibilities of scraping. However, scrape place by place would take a long time and was beyond the possibilities of the limits imposed by the platform. At the same time, it is thought in this way it is possible to partially balance the targeting of people interested in visiting this place.

Once the table with the data was obtained, we moved on to their processing, starting with a manual classification of the activities, to simplify a further step carried out later. That is the cleaning of text the posts descriptions to make them more easily classifiable for the construction of the final model knowing already which post was unclassifiable by description.

At this point, various outputs, graphics, maps, and diagrams have been built using programs such as GEPHI that allow you to do a network analysis.

3.1 Data extraction

Instagram no longer offers a native API for obtaining data from its platform. The policies do not allow the saving of personal data, so the possibility of obtaining information is limited. It was necessary to create a methodology that both respected the rules imposed by Instagram, but at the same time allowed to have a database that could be used for research purposes.

The choice was to write the code from the beginning. Due to the new restrictions after several searches in the bibliographies of previous works and search for packages ready for scraping, it was found that one of the most used methods requires scraping data through a browser. On the web different GitHub repositories show how to perform the extraction (e.g. see paragraph 3.1), and to get that it is necessary to use external packages. Among those identified, the most used were BeatifulSoap or Selenium, which were applied individually or combined with each other. Both allow the code to simulate the user actions, opening a browser that executes commands through the code. Internet sites usually transmit the information to be displayed graphically to the browser in HTML, and the package allows the user to query the code to obtain the information that the website sends to the browser.

The idea was not to save any data as required by the policies on the use of the Instagram platform, but to obtain only information made publicly available by users and the platform, such as links.

The work done in Python coding was divided for the scraping part into different parts: one to obtain information for each post that is realistic to the search hashtag and then to obtain the coordinates of the places where the posts were published.

The programming took some time starting from correcting the errors or avoiding them from happening during the scraping phase, the scraping itself took a long time and is more, the more posts to pass. This among the many was one of the reasons that led to analyse an area that did not have too high several posts, but which at the same time were statistically significant.

3.1.1 First part: getting the posts' link

Instagram allows you to search for posts related to a hashtag in a sequence of images that can then be clicked and viewed through a pop-up and directly from it also be queried by scrolling through the various posts.

Initially, it was thought to combine the first and second scaping parts together to reduce the time. However, Instagram limits user activity to a certain number of requests, and therefore it is not possible to make too many. In fact, the initial program simulated the user in switching from one post to another related to the hashtag #valborbera, but after several requests, the program then at some point crashed and could no longer go forward in the scrolling of the posts and of course once Instagram allowed you to be able to browse it again, it was necessary to start from the beginning.

This is because it was impossible to find an effective method to maintain a point of reference for the last investigated post. So, it was decided to divide this phase in two: the first where the idea explained above was maintained only to obtain the links of each post, while in the second part each post was loaded up to the possible limit of requests for construction of the dataset.

At this point, it was possible to switch from one post to another by slowly loading the entire list of images related to the hashtag and then scraping all the post's links. After that, it was easy to go on for the second phase, having available the complete list of links to upload. The advantage in this case was derived from the possibility of starting over from the last post after the interruption by the platform.

3.1.2 Second part: scraping the posts by hashtag

Once all the links have been obtained, the program could load each link and collect the information that was made publicly available by users and without saving any personal data of the user, but post after post, recorded:

- the publication date,
- the post's description,
- the location's link if present in the post,
- the post's link,
- the user's link

To make the requests, the Selenium package is used, capable of reading the HTMLcode and returning the value of a specific field within the web page. An item was added for every post writing some code to get the information using Selenium for each previous category. Precisely for this reason, this phase did not only take a long time to exceed the number of requests imposed

by the social network; indeed, it took several days to be completed, but it was also essential for each post to be sure an item would be appended in a list for each category, even if nothing was present. This is particularly important when building the dataset using the Pandas package. This extension allows you to load a list of lists relatively easily to obtain a data frame immediately. This is done by assuming each list within the list has the same size and structure.

Each list in the list was a row referring to a single post containing the previously listed categories. These categories in the final data frame will form the columns, while each row will correspond to a post.

The publication date is used to give a temporal attribute, thanks to it, it will be possible to query the data frame based on periods. Instead, the description can help characterize the content that the user intended to transmit through his post. The localization will be used in the next step for a further scraping of the coordinates to georeference the posts and therefore give a place to the content of the post (not always available), the link of the post is used to give an index to the data frame, each link in fact refers to a single content in Instagram, finally the user's link is useful to verify how interactions occur within the analysis area, for example if a user is of the place or not. The result contains:

- Links,
- Publication date,
- Posts description,
- Geotag (location link),
- Users' link.

3.1.3 Third part: getting the location coordinates

The purpose of the third coding part was to be able to georeference data, and to do this it was necessary to have coordinates associated with each post. They are contained in the page of posts associated with a place, that is, when the user searches for the name of a place, he will find the list of all posts for the name of the place he is looking for. To get to the coordinates, it was therefore necessary to load the page relating to the place and through the HTMLcode try to obtain the coordinates. To achieve this, the procedure was the same as in phase two, having already obtained the links of the locations entered by users (Table 1).

It was possible to have the repetition of places and, therefore, to prevent the program from loading more than once the same were duplicates.

Table 1. Third scraping table result (geotag links and WGS84 coordinates)

Geotag	lat	lng
https://www.instagram.com/explore/locations/777374264/bar-tre-scudi-borghetto-di-borbera/	44.730	8.945
https://www.instagram.com/explore/locations/164544790222568/monte-carmo/	44.615	9.198
https://www.instagram.com/explore/locations/270991126352171/torrente-borbera/	44.725	8.953
https://www.instagram.com/explore/locations/319769615/sport-hotel-prodongo/	44.695	9.26
https://www.instagram.com/explore/locations/1347264935407707/val-borbera/	44.726	9.030
https://www.instagram.com/explore/locations/55689/maialino/	40.738	-73.985
https://www.instagram.com/explore/locations/591716044255427/daglio/	44.644	9.149
https://www.instagram.com/explore/locations/107904602605018/monte-giarolo/	44.727	9.128

https://www.instagram.com/explore/locations/416200764/montemarzino/	44.848	8.991
https://www.instagram.com/explore/locations/754358562/ristorante-costata/	44.400	8.957
https://www.instagram.com/explore/locations/655568287794551/dova-superiore/	44.385	9.030
https://www.instagram.com/explore/locations/298198903697317/cosola-arc/	44.669	9.178

3.1.4 Fourth part: get where the user usually posts

The last step of scraping was to upload each user link and see where they usually post the most recent links. The program then used the initial idea of opening the first post on the page and then moving on to the next trying to get the information requested, the location. The code after ten locations were found was passed to the next user, in the same way after it failed 10 times. Therefore, there may be users who have fewer than 10 locations found.

3.2 Pre-processing and data cleaning

Once the scraping phases have been completed, the concrete analysis part started, but initially a step was still necessary to carry out for data cleaning and rearrangement processes.

For the classification of the data two models were built (both based on the description of the posts, where there were no posts were discarded): a model for the classification of activities, another model to bring all the locations to a point reference.

As previously mentioned, the package available in Python used to build a model was Sklearn, which needs to clean up the text to increase the accuracy of the prediction to make the words easily recognizable.

3.2.1 Pre-processing for model creation to predict the activities

To increase the accuracy of the activities, in a separate table with only the links column and the description of the posts, we started from a manual classification of the activities carried out for each post, since the classification was done manually, and it was easier to identify the non-classifiable posts. In fact, a classification was made that also included a “Not classifiable” class to facilitate the construction of the next model to bring all the places back to points.

The classification was initially conceived, also through a vision of the description of the initial posts in:

- Community, events that do not fit into the other groups following such as meetings with friends, political discussions, etc.
- Event, contains posts that talk about village festivals, events that also attract people from outside for music concerts, visits to museums or group walks in the nature.
- Family, events that concern the family such as weddings, communions, meetings between relatives, etc.
- Food, relate to all posts that talk about food and can refer to cafes, restaurants, or places where local food is tasted.
- History, posts that contain photos with historical monuments in the area and historical stories about some event that took place.
- Nature, concerns all posts that contain images with a landscape, speak of the surrounding environment or times of the day.

- Not Classifiable, it includes all posts that do not contain any description or a few unclassifiable words.
- Religion, all posts about religious events or about religious monuments.
- Shopping, post where we talk about purchases.
- Sports, all posts dealing with physical activity such as trekking, running, walking are collected
- Work, contains all the posts where we talk about work, in this case we talk about agricultural activities.

These classes were then revised later and unified during the construction of the charts.

3.2.2 Pre-processing for model creation to predict the locations

As told already in the paragraph before, the first step was to make it as easy as possible for the Sklearn module to read the words and bring the classification of all the locations back to points. To use Sklearn to make classifications from the text pre-processing is necessary to clean characters that can reduce the final accuracy, together with what has already been done previously to identify non-classifiable posts. The processing was applied to the text of the descriptions of each post collected during the second stage of the scraping.

Therefore, an attempt was made to identify all the localities that cannot be referred to as the punctual dimension or in any case close to it. The simplest step was to eliminate the locations of the municipalities, an automated process thanks to the use of the Jupyter Notebook. This was done because by having the names of the municipalities from the shapefiles on their borders, it is possible to report the name in the same format as the location of Instagram. For example, on Instagram it appears as “cantalupo-ligure”, while the real norm is “Cantalupo Ligure”. Through the code, therefore, it was brought back to the same format by replacing the space with the bar and at the end deleted all the locations related to the municipalities and “val-borbera” to get the result before mentioned.

Then some elaborations were carried out on the text to split words. Several tests have been done and different methodologies have been extracted from websites and the bibliographies, in the end a mediation was made between the two parties. Among the most effective was an article published on the Stack abuse website, but it was still particularly difficult how to classify the hashtags as they contained several non-divisible words because there was no character to distinguish the beginning or end such as a space, comma, etc.

Among them the most relevant was dividing each word in the row corresponding to the description of a post into different elements, together with eliminating other elements such as spelling characters. Then each word has been split as an element of a list so that the program performs the classification based on the occurrence of words for each line. The subdivision also made it possible to detect the hashtags, usually more present in the description than the simple text. This step was particularly important to eliminate the hashtag used to search for posts (#valborbera), which was naturally present in each of them.

Another step for the analysis was cleaning the post description column. First, we tried to remove all the special characters contained in the posts. It is common to have emoticons that do not describe what is happening. All the single letter or number characters within the text were then

eliminated, just to try to classify only the words. Finally, the letters were changed to lowercase so that all the same words were the same even for Sklearn.

3.2.3 Increase in the area considered using a buffer, select only locations within it and delete duplicates

During the scraping and pre-processing phase, an error of even a few meters was noted in the localization of some places outside the boundaries of the valley, it was therefore decided to create a buffer also to maintain some external localities with which the study area has important exchanges.

Having as interest only what happens within the study area, at the beginning it was decided to select only the locations that fell within it, together with the related posts. However, it was noted in this phase that some of the localities also influenced the movements outside. Although about 40% of the places were inside the buffer, the majority, more than 90% of the posts were linked to those places. This shows that in general users had a good knowledge of what the actual boundaries of the valley were by correctly using the hashtag #valborbera.

It was noted the posts outside the valley were far fewer in number than the internal ones, and many of them were made by people who usually publish inside, who probably publish outside to identify their area of origin.

Some geotags have the same name, but different links, always referring to the same place. So, to simplify the work, we try to have the same coordinates for the same geotag even if the link is different. In addition, some of the locations have an incorrect location, a problem also detected on Instagram and therefore safe. This could be one of the reasons why there are duplicates of the same location. For this it may be possible to add a further selection of coordinates within the area to choose the exact location.

3.3 Processing and data analysis

Once the dataset was prepared, the next step was the data processing. Anyway, for the improvement of the dataset, we returned several times to the previous paragraph to improve ever more accuracy with which the data were classified.

In fact, in this part of the process, an attempt was made to classify the activities carried out, through the manual construction of a dataset to make the model recognize the data that could be classified, and which could not. After that, we moved on to the direct classification of the places, trying to solve the problem of how the localities could refer differently to different areal places.

3.3.1 Classify the activity made in the post

The classified activities are consistent only for some sections, particularly Nature and Not Classifiable, in a smaller number of Sport and Food. The others are not consistent, and therefore it was decided to bring them back to a single class.

For the classification of the activities carried out within a post, a sample of posts was manually classified. For this purpose, a table containing two fields was initially exported: one with the description of the post to understand what the purpose of the publication was, while the other field the unique links for the identification of each post to lead back to the original table by joining the classified data.

To classify the activities carried out for each post, it was initially necessary to build the field by manually classifying some posts on which to build the model. The manual classification was done on around 200 posts to label all the others. The last three lines show: 1) accuracy, how many elements of the Sample for the construction of the model have been correctly classified; 2) macro average (Macro avg), the average of the unweighted mean per label; 3) weighted average (Weighted avg), the width of the weighted average of the media per label (Table 2).

To understand which model was used and which algorithm was applied, see the next paragraph (3.3.2). The same steps were also used to classify activities.

Table 2. Accuracy report for the activities' classification

	Precision	Recall	f1-score	Support
Community	0.00	0.00	0.00	8
Event	1.00	0.20	0.33	10
Food	0.72	0.67	0.69	42
History	1.00	0.20	0.33	5
Nature	0.68	0.88	0.77	93
Not Classificable	0.74	1.00	0.85	40
Religion	0.00	0.01	0.02	1
Sport	0.80	0.65	0.71	54
Work	1.00	0.54	0.70	13
Accuracy	-	-	0.73	267
Macro avg	0.59	0.41	0.44	267
Weighted avg	0.73	0.73	0.70	267

3.3.2 Predict the location where missing and solve problems on the geotag

As explained in the previous paragraphs, not all posts had a location reported. However, it was indicated in the description of the image. To extract it automatically it was necessary to build a model. The process is explained below.

With this methodology, two different columns were created, one with the entire description extrapolated from the posts and another with only the hashtags to check which of the two gave the best accuracy. In addition to this, Sklearn allows you to apply different types of algorithms, at the same time, combine them or decide which prediction to choose between different calculations based on a pre-established weight.

Among the algorithms tested the ones with the best accuracy were: MultinomialNB, SGDClassifier, LogisticRegression, SVM and RandomForestClassifier. the first two had an accuracy below 50% while it was above 50% for the last three. As showed on the paragraph 3.2.2. several attempts have also been made between using the full description or just hashtags, and in most cases the use of the full description has been more accurate. This is because some posts have no hahstags and therefore removing the description the field remains empty.

Furthermore, the localities present referred to different geometric entities, both punctual and areal (for examples, economic activities such as restaurants or municipalities). Therefore, we have tried to take only the punctual localities and predict specific places also for the areal places. Then there was also an attempt to eliminate hashtags related to municipalities, but in this case,

the accuracy decreased. The solution was to keep only the locations on point and keep the hashtags except of course, the one used to search for #valborbera posts.

In this way, despite the accuracy above 50%, about one-third of the total posts remained with the punctual location. So, 50% increased the total to two-thirds.

Finally, through use, one of the problems solved is that of the different dimensionality of the geotags. In fact, in some cases it could refer to both an area and a specific place. An attempt was therefore made to redefine the classification of geotags to places as punctual as possible.

3.3.3 Splitting between locals and not-locals

To check if a user we published is local or non-local, two methods were used: one of them was verifying how many times the same user posted as indicator as also assumed in the paper "Using Flickr Geotagged Photos to Estimate Visitor Trajectories in World Heritage Cities" (Domènech et al., 2020). In fact, it is more likely that a person living in the place is published more, while a person outside the analysis area posts little or only a post. The other went user by user looking for which locations he indicated in the posts of his profile.

However, some users were no longer viewable, and others had become private. This could be due to several events that have occurred, including: the cancellation of the account, the user has decided to change the view of their profile from public to private, etc. In this case, the possibility to indicate whether a user was local or not referred only to the number of posts published.

If the user had published a significant number of posts within the area he was considered "local", if in the surrounding area as "around", outside "non-local". Furthermore, the definition was also based on the number of total publications. In fact, it is customary to think that a user who publishes several times is local, while on the contrary not.

3.3.4 New locations found

From the scraping of the profiles that published mentioning the hashtag #valborbera, new things became known, including new places that had not previously been detected by locations in the posts (Figure 5). Most of the places not previously detected through the hashtag #valborbera are located outside the borders of the Borbera Valley, especially on the valley floor and in the most populated town to the west. As for the new places within the Val Borbera, they are above all near the inhabited centers.

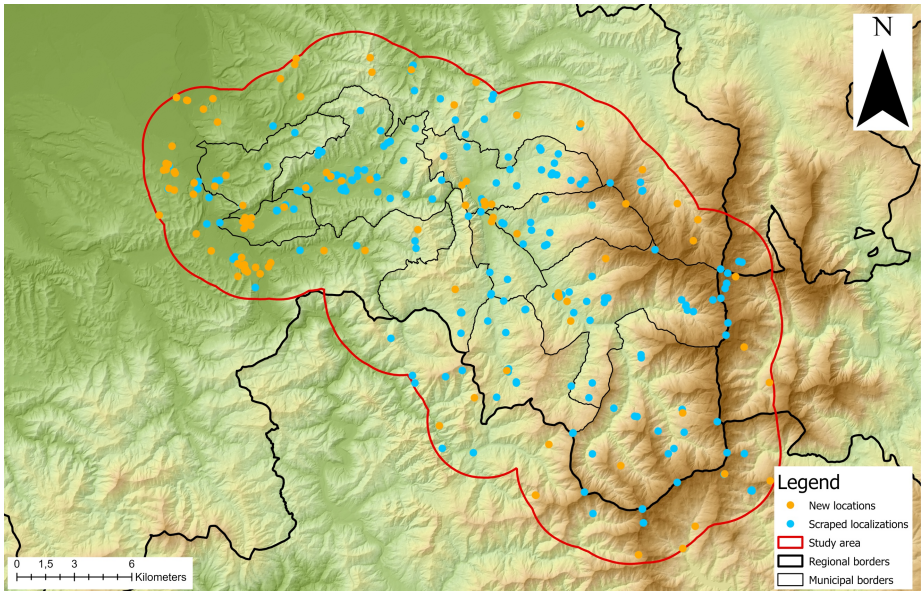

Figure 5. Locations found through the hashtag #valborbera and through profile analysis.

Among the various new places not previously detected by the scraping of the hashtag #valborbera there are places with an important impact on the study area. However, they are not reported because the purposes for which these places are frequented do not directly concern people who have an interest in reporting the context in which they are (Table 3).

One of the cases is the locality indicated in the table as “mulino-club”, a discotheque, an important meeting point of the area, in the immediate vicinity there is also a pizzeria and hotel (Il Mulino), a restaurant (RaviolPub) and an aquatic park (BolleBlu). All other places are identified by the hashtag scraping, but not consistently as are the publications for the respective on Instagram.

It is different from most of the other places that should be closely linked to the context in which they are located such as the “castello-degli-spinola-di-montessoro”, a castle, “comunita-montana-valle-borbera-e-valle-spinti”.
At this point it would be interesting to try to investigate each location to verify the frequency with which a geotag referring to it is mentioned, but the limitations imposed by Instagram would take a long time to achieve the result. The analysis of the localities would also be useful for the validation of what has been learned only through the hashtag.

Table 3. List of new locations obtained from profiles

Location name	latitude	longitude
50-special	44.728	8.970
albera-ligure	44.703	9.067
artinvalle	44.594	9.135
autoborbera-sas	44.715	8.913
bar-al-vecchio-san	44.725	8.868
bb-il-convitto	44.674	9.094
borghetto-di-borbera	44.730	8.944

cabella-ligure	44.674	9.090
campo-sportivo-vignole-borbera	44.709	8.886
cantalupo-ligure	44.716	9.033
carrega-ligure	44.619	9.176
castello-degli-spinola-di-montessoro	44.707	8.891
comunita-montana-valle-borbera-e-valle-spinti	44.716	9.046
croce-rossa-italiana-comitato-di-vignole-borbera	44.705	8.887
cuore-di-pane-bio	44.675	9.094
hotel-ristorante-da-bruno	44.710	9.051
il-casone-di-rosano	44.662	9.102
la-stalla-dei-ciuchi	44.717	9.050
maura-coiffeur	44.717	9.046
mongiardino-ligure	44.639	9.060
monte-spineto	44.716	8.886
mulino-club	44.724	8.927
non-ti-scordar-di-me	44.718	9.045
parco-mongiardino	44.673	9.095
pobbio-superiore	44.691	9.125
roccaforte-ligure	44.677	9.020
rocchetta-ligure	44.709	9.051
selvagnassi	44.671	9.100
stazzano	44.730	8.875
tenuta-basini	44.708	8.889
variano-inferiore	44.728	8.951
vignole-borbera	44.710	8.892

3.3.5 Finding clusters and hotsposts using Network Analysis

The first step was to georeference the posts using the coordinates obtained from the locations indicated in the posts to see how many were made within the area of interest and how many were not. It was found that most of the downloaded places are outside the study area and that at the same time, selecting only the actual municipalities of the valley, hotspots that belonged to the cluster of people moving from the valley were not included. For this reason, a buffer was made around the municipalities to include those areas as well.

Many of the bibliographies use programs to interpret how people move within the study area through maps, charts, and diagrams using different programs. Among them, the most popular and most often mentioned is GEPHI.

At the end of the data processing, it was necessary to build the tables in a format suitable for the understanding of GEPHI. A table must contain the nodes of our network with the identification codes for each of them and a weight. The identification code will be used in the second table that must be created for the edges that indicate the attributes between the various nodes, for example, it is possible to know how many people have passed from one node to another.

The edges table must contain the identification code from which the movement starts (Source) and to which it arrives (Target). In this case, we do not know if the user has made this move, but he did it in the analyzed period. Therefore, for the construction of the edges table, it was necessary to verify where he started from and where he arrived in the study period and to add

these equal movements between the different users to verify the actual flow between the users between the different nodes.

To arrive at a solution, it was decided to build the nodes on the locations indicated in the posts, while for the construction of the table on the edges check the people who had visited more than one place. The problem is that within the diagram, there will be no data regarding users who have published only one post for obvious reasons, since they do not have more than one place, it is not possible to create a network by the user. Then a list is created for all users and a list for each place.

The process of building the tables required many steps. First, the list for the place names was created, and another with the usernames. First, all the places visited by a user over the period are identified. After it is checked that there is more than one post for the place because it would not make sense to show a place with only one visit, it would make reading the final diagram even more complicated. At the end for each user the couple of places visited is added to the list which will then be transformed into a Pandas data frame, checking that the same coupled is not created.

At the end, the Source and Target columns are created by joining the names to have the identification code for both the columns of the place of departure and arrival.

Other information is added to characterize the final diagram, such as: the name of the place, the month of greatest popularity for each node, the number of visits per month and years, etc. Furthermore, the program allows you to do some statistical calculations automatically, for example to see the level of data clustering.

4 Results

This chapter aims to show the results achieved through all the processes previously mentioned. We will describe the general statistics, such as the number of posts analysed and how many were classifiable.

Following, those for a better understanding of the data as is the distribution of the publication of posts during the year and the years taken into consideration, which are the most popular places and which activities are carried out in that place and check if it is possible to recognize people who publish local or not and if they have different trends.

Finally, we will try to recognize if there are hotspots to be recognized within the area.

4.1 General statistic from the scraping

Instagram gives the total number of posts that have used the searched hashtag, but obviously, when the page is open only shows those of public accounts. For the hashtag #valbobera, most of the posts linked to it are public, in fact it was possible, out of a total of almost 24,000 posts, to obtain information of 18,651 posts (Table 4).

The number of locations indicated in the posts is more than 650. However, after the process of filtering the posts cutting out all those who did not fit into the study area and using the model created with Sklearn to bring all the locations back to point level, the remaining are greatly reduced to about 240.

Despite the clear diminish of the locations inside the study area, there are still many posts classified, circa 11000. Slightly more than half of the posts have an added localization, and out of a total of about 8700 inhabitants, the users who publish inside the area are about 1772; this number is also reduced after the final processing from the 2700 users of the beginning.

Table 4. General statistics on processed posts

Total posts related to the hashtag #valborbera	23.978
Total posts scraped	18.651
Total localities scraped	656
Total localities inside the study area buffer	237
Posts filtered after classification by model	11.247
Total users publishing in the study area	1.772

4.1.1 Trend of the #valbobera hashtag in the years and months

The popularity of the hashtag #valborbera has grown almost exponentially since 2012, when the first post was published. After that, it began to be constantly popular starting from a small number and becoming consistent only in recent years (Figure 6). From the first year in which the hashtag began to be used with a total of 30, today there are more than 3000. There is a substantial jump between 2019 and 2020, the Covid period could have influenced the trend. Statistically, the years 2012, 2013 and 2014 have little weight compared to all the others, and it will make more sense to analyse the data by aggregating them monthly.

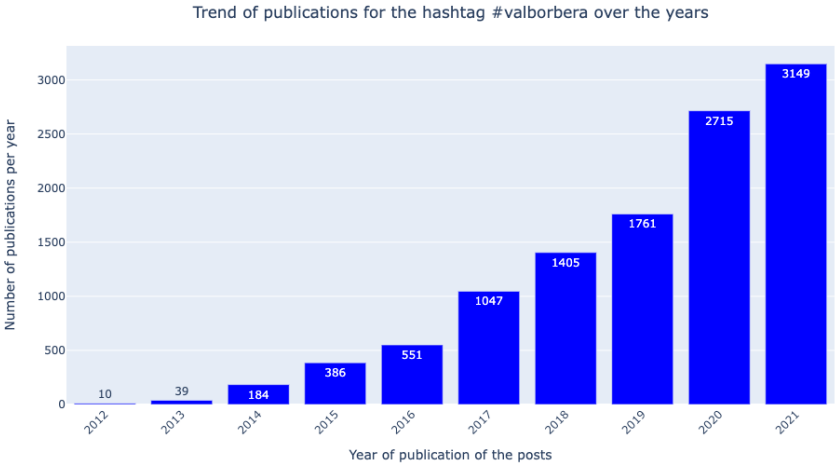

Figure 6. Trend of the publications for the hashtag #valborbera over the years

But before looking at the aggregate data, let's see the different trends over the years month by month (Figure 7). It is interesting to note how generally the trends are similar monthly every year, above all starting from 2017 (year in which the hashtag begins to be statistically consistent) with a trend that turns out to be similar every year: in the period of late spring and summer (May-August) the publications almost increase compared to the rest of the year. The last two years seem to be an exception to the trends of previous years for many of the months, which have instead grown steadily. For these two years, the months of January, February, March, May, June, July, August, October, and November are particularly noteworthy.

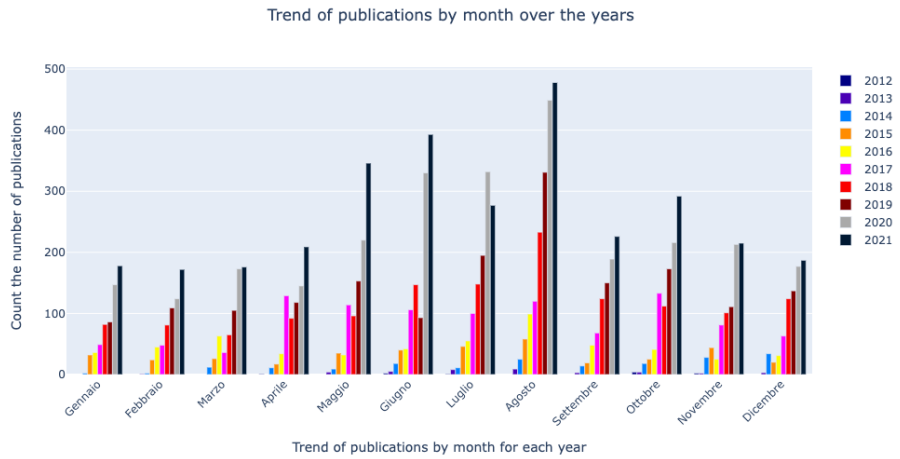

Figure 7. Trend posts publications by month over the years

In general, it is said that there is a clear popularity of the hashtag #valborbera in the summer period with an important peak in August. in fact, the publications generally for the other seasons of the year, considering the trend year by year, remain stable. This is especially true after 2017. August represents an important exception compared to the other months, extremely true for the other years outside of 2020 and 2021. The increase is significant for the other years while for 2020 and 2021 the margin of posts published between August and the other summer months is reduced.

4.1.2 Trend number of posts published per user

The following graph shows how many users have published one, two, three posts and so on (Figure 8). Immediately it comes to the eye how a high number of users have published less than 200 posts and there are some exceptions of users who have exceeded this threshold some of which significantly. In any case, there is a consistent group of users who have published below 4/5 posts. People who have published only one post in the time span analysed have a specific target of places visited, which are mainly points of particular interest from a gastronomic point of view or for sports activities.

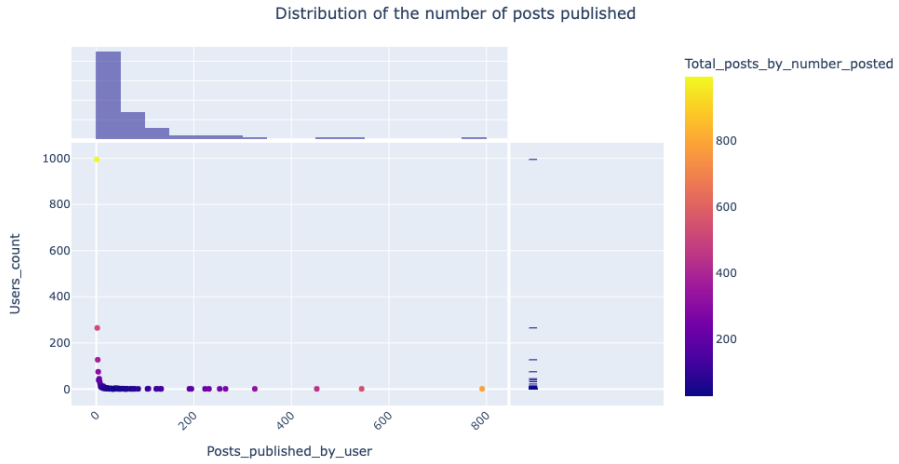

Figure 8. Number of posts published for each user

4.1.3 Frequency of the publications per day

Having the opportunity to know the publication date of each post (it would also have been possible to obtain the time, but it was not particularly useful in this work, so just day, month and year was got), it was possible to develop a graph in “stock exchange” style (Figure 9).

It is notable how there is a general trend that leads to an increase every year in the number of posts that use the hashtag #valborbera and at the same time similar annual trends. For example, there are peaks in the middle of the year, which are characteristic of the month of August by going to see the graph specifically divided by months and years.

As seen in chapter two of the classification, the territory has undergone a strong displacement which has led to much of the elderly population, less dedicated to the use of social media.

Furthermore, in the time interval considered, only the last few years have a relatively important consistency with a growth of daily posts precisely in these; however, there is a difference between summer and winter.

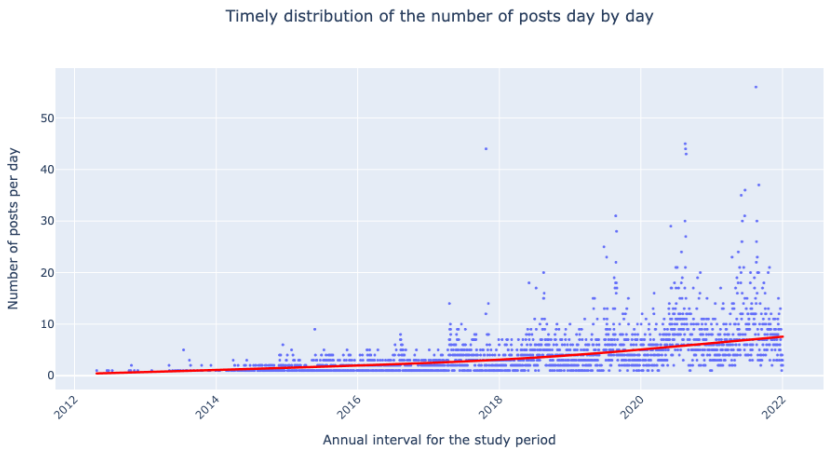

Figure 9. Time-series of post publication day by day

4.1.4 Most popular activities

Among the most popular activities found, those that fall into the nature category clearly impose themselves on the others with a trend that goes to a minimum in the winter period to have a peak in the summer period, especially in August where it reaches more than three times the number of posts published in the minimum period.

Considering what has been said in the previous paragraphs, where it has been seen predominantly how the area is frequented especially in the summer period and how some meeting places and communities such as the “Il Mulino” disco are missing, it is likely that the hashtag #valborbera is the expression of a specific target of users interested in outdoor activities.

Anyway, the activities that fall within the sports category have a generally stable trend for all seasons and months, although some peaks can be recognized especially at the beginning of the hot period of the year in May and February. On the other hand, in the food category, it is possible to recognize a trend during the year with lows at the beginning of the cold October period and at the beginning of the hot period while the maximums in early summer (June) February.

All other categories have a minimal impact on the overall trend, but the category of unclassified posts remains which has a major impact. Eventually these categories will be aggregated for later analysis.

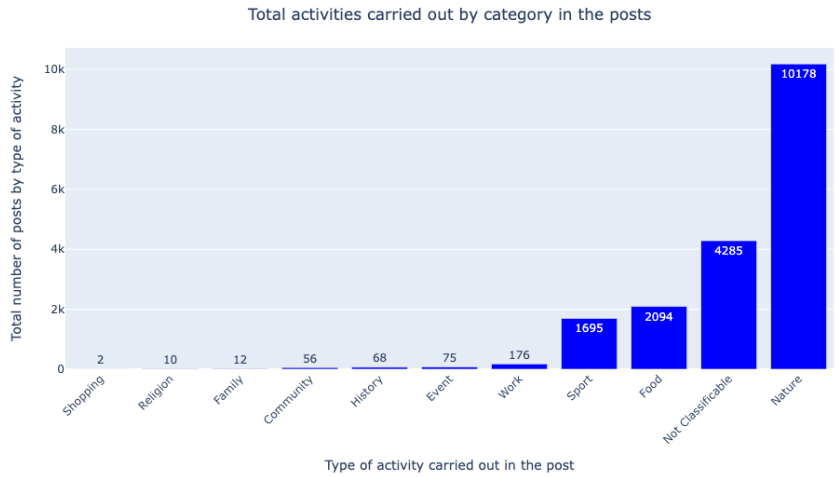

Figure 10. Total activities carried out by category in the posts

The following graph (Figure 11) shows the aggregation of the other types of activities previously seen (Figure 10), which do not have a particular trend during the year, except for the categories: Food, Nature, Sport. The aggregation was made for all the years together but divided monthly because as already mentioned, some of the years do not have a particular statistical significance compared to the total of posts published during that year.

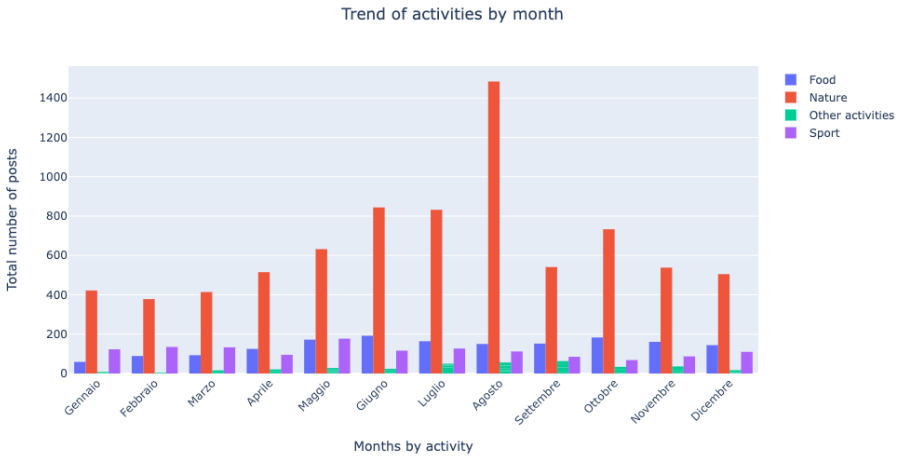

Figure 11. Trend by month for each category type

It is interesting to see the trend for the nature category, which remains far above the numbers of the other categories for all months, starting from the minimum of January and February and then starting to rise in March and stabilizing in June and July, but then with a sudden peak in August and fall sharply again from September until returning to January.

The numbers for the Sport and Food categories remain very low throughout the year but it is still possible to identify a trend. As regards Food in particular, there are two peaks in June and October, almost opposite to the Nature Trend. The Sport category, on the other hand, seems to have a more irregular trend, this is because it could be influenced by different factors such as the climate.

4.1.5 Most popular hashtags

To have a quick validation of the main activities carried out, it is possible to give a look on which hashtags are the most popular (Figure 12). In fact, many times the user inserts a hashtag in the description of the post connected to the activity he is carrying out. You can visualize how among the most mentioned hashtags there are those that contain themes about nature, the landscape, the environment in which the person interacts.

To identify the hashtags concerning food, however, you must go below the number of 400 per hashtag, while there is a more convincing confirmation regarding the sports category in fact they are identified among the most mentioned: #trekking and #trailrunning.

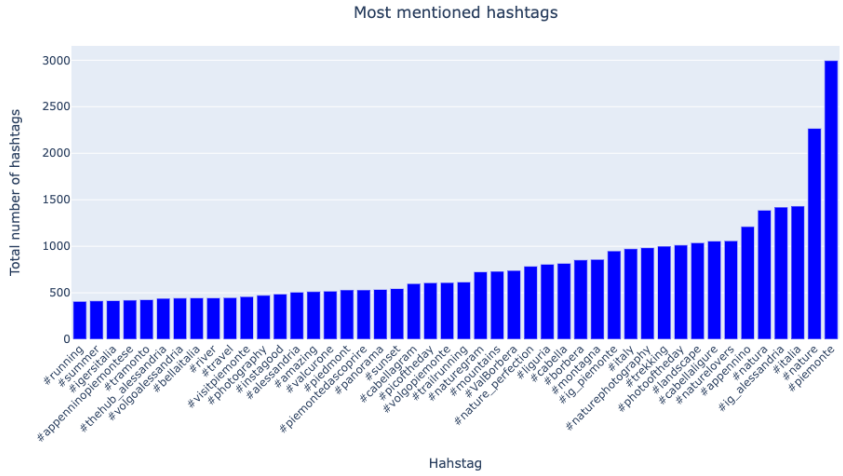

Figure 12. Most mentioned hashtags

4.1.6 Most popular places

The following graph (Figure 13) generally shows which are the most popular places, later a differentiation will also be made on a seasonal level and between local and non-local. Considering the top 10 most visited places, 5 are natural places, so you can see how most posts refer to places contextualized outdoors.

Among these ten, four of them refer to the same place, one is the Borbera river, the other three are specific to the Strette di Pertuso area (Boscopiano and Locanda Pertuso). The Borbera is the river that flows throughout the valley and the Strette di Pertuso is a part of this river where there is also Boscopiano, an area equipped for picnics from which it is possible to reach the river and more above by means of stairs. the Locanda Pertuso is located.

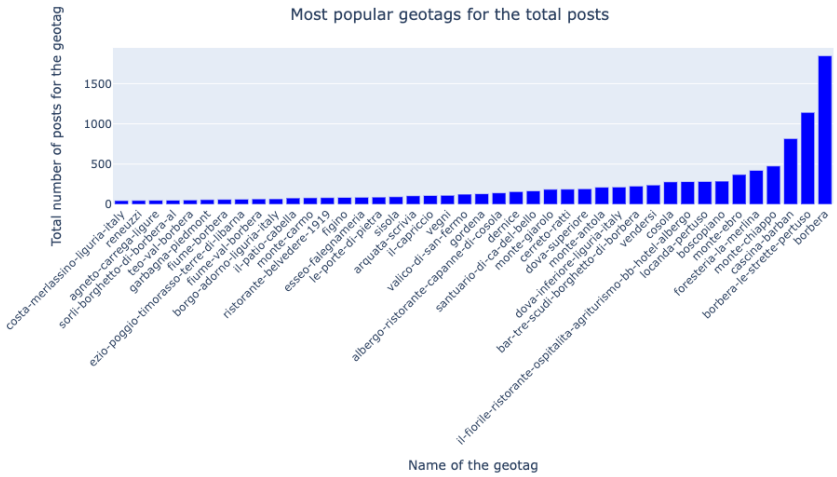

Figure 13. Most popular geotags for the total posts

4.1.7 Most popular places by year

The first year 2012 (Figure 14), we have already seen not to be statistically significant, in any case it is possible to see which places have started to be mentioned. In this first year of the 6 I mentioned 3 were related to natural areas, two to restaurants and one to a locality. However, the Borbera river area already has the largest number of references.

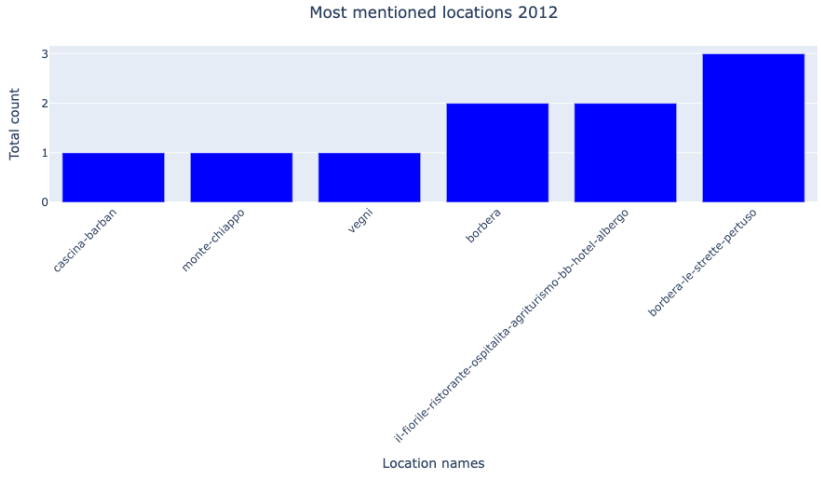

Figure 14. Most mentioned locations 2012

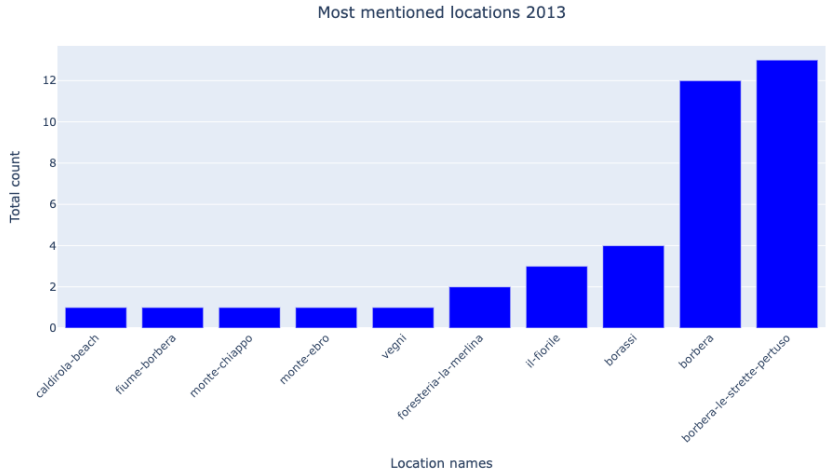

Figure 15. Most mentioned locations 2013

The second year is not statistically significant again but compared to the previous one (Figure 15), we can see a significant increase in the number of posts and how many have mentioned the Borbera river and its Strette di Pertuso area.

The 2014 begins to be a statistically significant year for the hashtag #valborbera (Figure 16), interesting that for this year the most mentioned location in the posts is nothing related to the Borbera river but the Monte Chiappo. Therefore, it would perhaps be interesting to verify if any event had happened in that locality in that year.

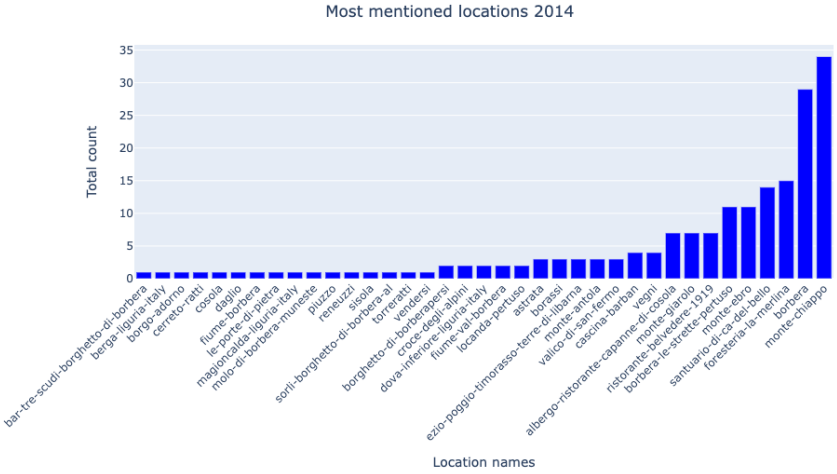

Figure 16. Most mentioned locations 2014

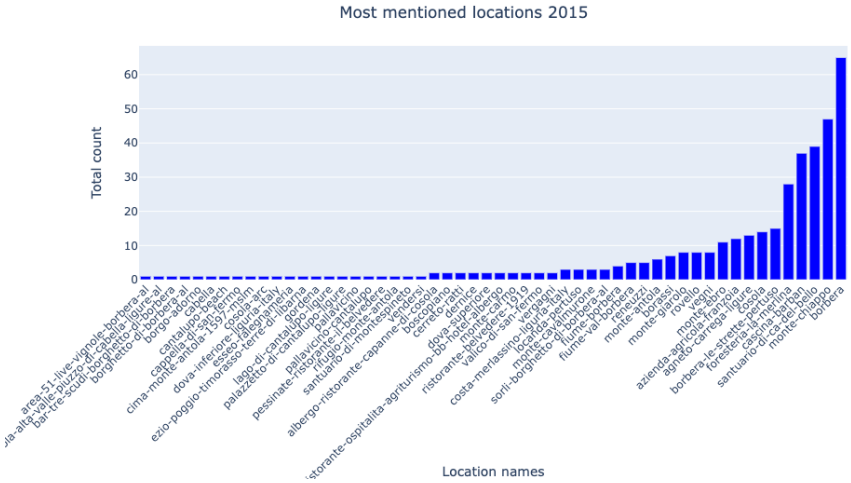

Figure 17. Most mentioned locations 2015

The 2015 (Figure 17) sees the return of the Borbera river among the main places, and at the same time the locations mentioned grow so much that it is necessary to make a cut in the places shown in the graph (greater than 1). During this year and the one before the locality of the sanctuary of Ca di Bello is among the main mentioned, while it will start to go down from 2016 (Figure 18), this could be due to some popular event that was held there every year and just after 2016 it started to be less followed.

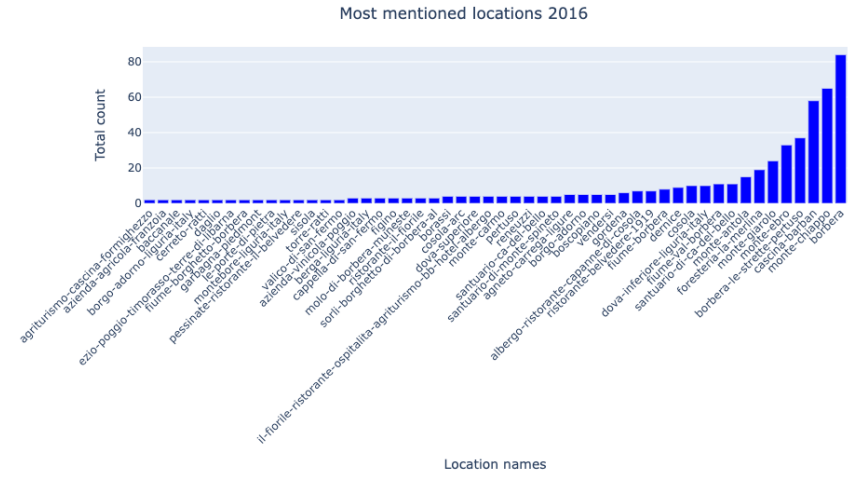

Figure 18. Most mentioned locations 2016

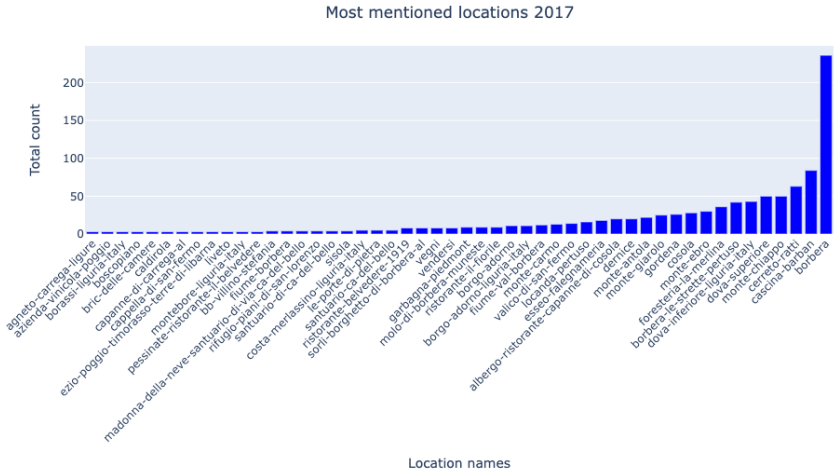

Figure 19. Most mentioned locations 2017

For 2017 there is a significant increase in the places recalling the Borbera river about 100 when the previous ones went just above 80 (Figure 19).

The same trend can be seen for 2018 (Figure 20).

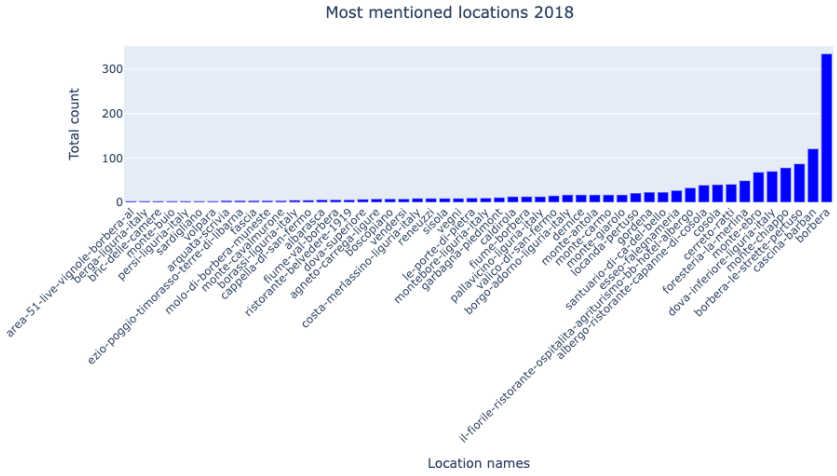

Figure 20. Most mentioned locations 2018

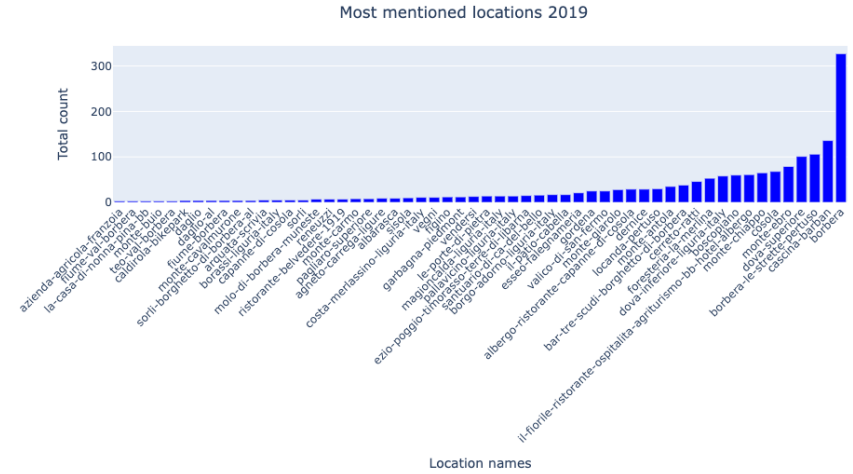

Figure 21. Most mentioned locations 2019

For 2019 there are no differences compared to 2017 and 2018 (Figure 21). The locality of Boscopiano begins to be numerically important during this year.

Starting in 2020, we are starting to see important numbers for a greater number of locations, reaching a significant number of posts per location (Figure 22). For example, the top 8 most mentioned places are mentioned by more than 100 people and the top 12 by about more than 50.

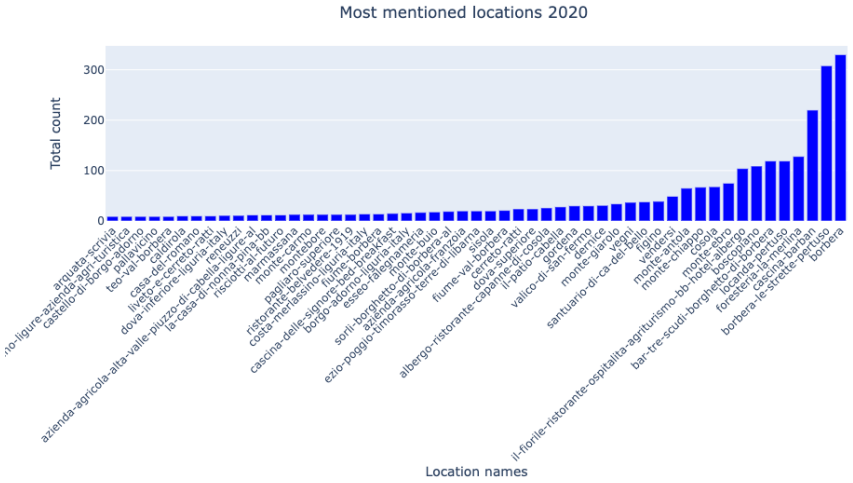

Figure 22. Most mentioned locations 2020

In 2021 the geotags referring to the Borbera river and the Strette di Pertuso reach an important number compared to all the other places around 400 (Figure 23).

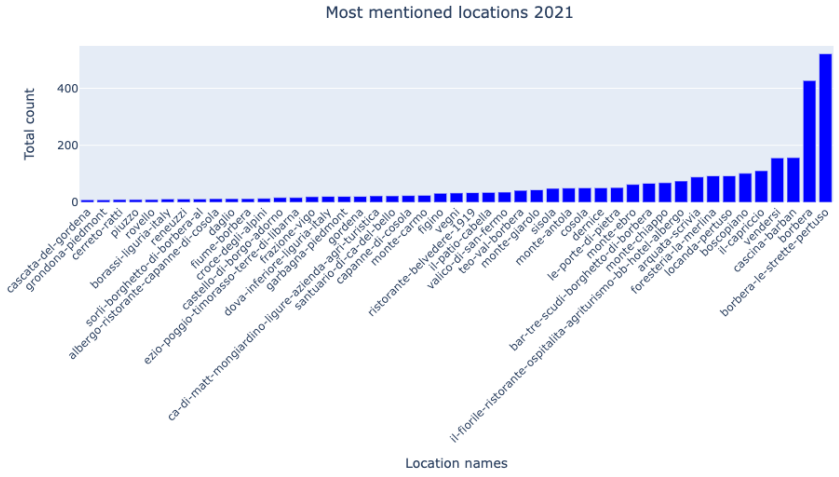

Figure 23. Most mentioned locations 2020

4.2 Local and non-local users

The following paragraphs show the data obtained from the definition of users between local and non-local. We initially started with a categorization divided into three classes:

- local, people who publish mainly within the study area.
- around, people who publish mainly in the provinces around the study area, including the one in which the area of interest falls.
- not locals, people from outside the provinces on the edge of the study area.

As already mentioned at the beginning of this chapter, having stated the general statistics on the scraping obtained, the total users found to have used the hashtag #valborbera are more than 2500 which is reduced to about 1800 considering the users who have published a post attributable to a place within the study area.

As we will see of these, the majority are non-local, even if the posts are published, they are shared on an almost equal percentage between the two parties.

4.2.1 Identify the origin of the users

Thanks to the work done to differentiate if a user comes directly from the study area, around it or far from it. It is possible to see that there are much more non-local users publishing, but local people's posts are numerous. On the other hand, the situation is particular for people who used to publish in the vicinity of the area, in this case, there is a borderline situation that is not totally definable, these users could in fact be partly attributable to locals or non-locals, they are numerous but publish little. However, it was decided to include them among the premises since for most cases they were people who had published more than once in the area and not in the same period.

From the Figure 24, non-local users are about 73% of the total while those coming from the surrounding provinces are about 19%, while those strictly local are about 8%. In total a 17% can be considered local.

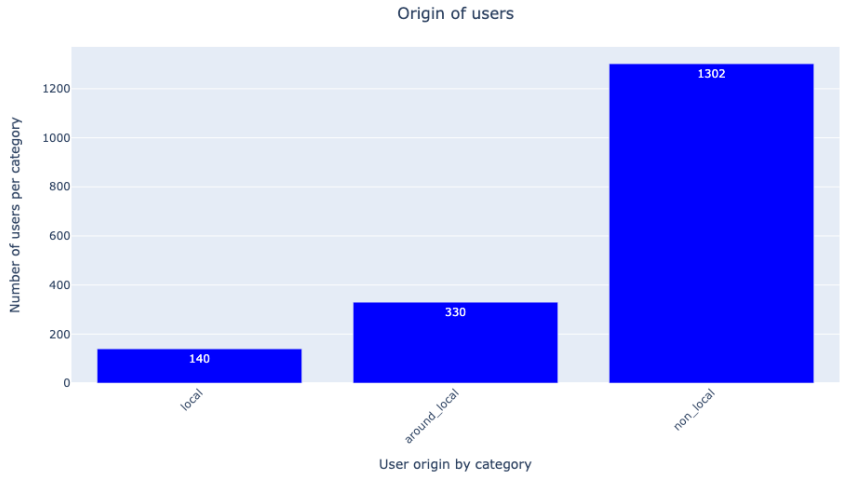

Figure 24. Origin of the users: local, from the study area, around, from the provincies around the sudy area, non-local from far to the provinces

It should be noted that the statistics almost match the number of posts published per user (Figure 25). In fact, 75% have published only one post, while about 25% have published more than one post. Of the latter, the largest share is 10% who published 2 posts, while 5% published 3 posts. This information will be taken into consideration for what we will see later.

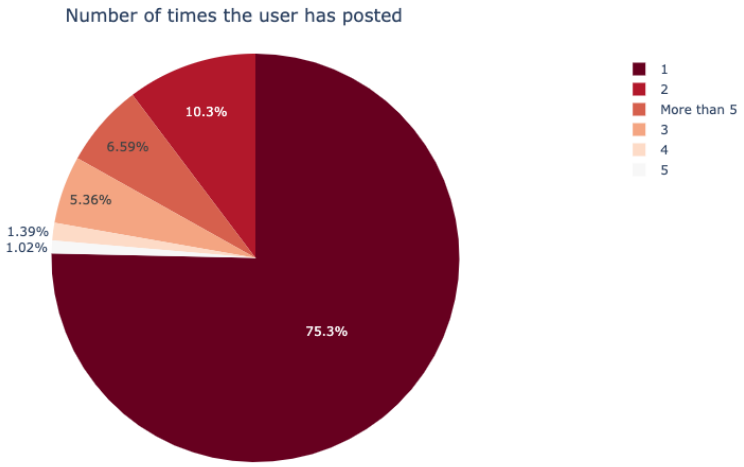

Figure 25. Number of times the user has posted

4.2.2 Trend of the #valbobera hashtag in the years and months in absolute values and not for local and non-locals

It notable a clear difference between the monthly trend over the years between people who usually publish in the territory of interest, compared to those who have published only one post or are external (Figure 26). In fact, in various ways it is possible to see how non-local people visit #valborbera, especially between June and August. Although the number of total users is just over half from the outside, but the premises have a much higher number. There is also a slight difference in the places visited, in fact, non-locals visit mostly restaurants or just particular points of natural interest (for example, mountain peaks).

As you would expect the locals being always present in the area, should have a greater consistency in the publication of posts. This happens for much of the year except for the summer period between May and August where the frequency increases with the peak in August.

Generally, an increasing and constant trend year by year is possible for all months, only 2020 and 2021 represent a strong exception for a few months, especially in April, May, June (2020 have one of the most important impacts compared to all the others), August and January. These differences could be due to the restrictions due to the covid, and for this reason they could be the object of further study.

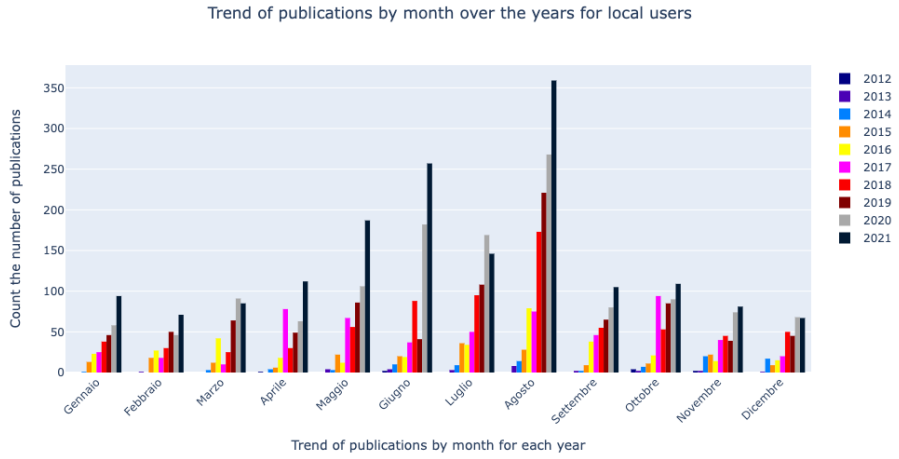

Figure 26. Trend of publications by month over the years for local users

Contrary to what was expected, however, the frequency of publications by non-locals seems to have a constant trend for half of the year and then grow strongly in the summer, but still partly returning to high levels in October (Figure 27).

There is a strong difference in the trend for the years 2020 and 2021 compared to the other years. This could also be due to the restrictions imposed during the Covid pandemic. In any case, the difference between 2020 and 2021 compared to other years is much more marked for non-locals when compared to the graph of the premises.

Another important difference concerns the total number of publications for the same month, those of non-locals are generally higher, almost double, than those of locals.

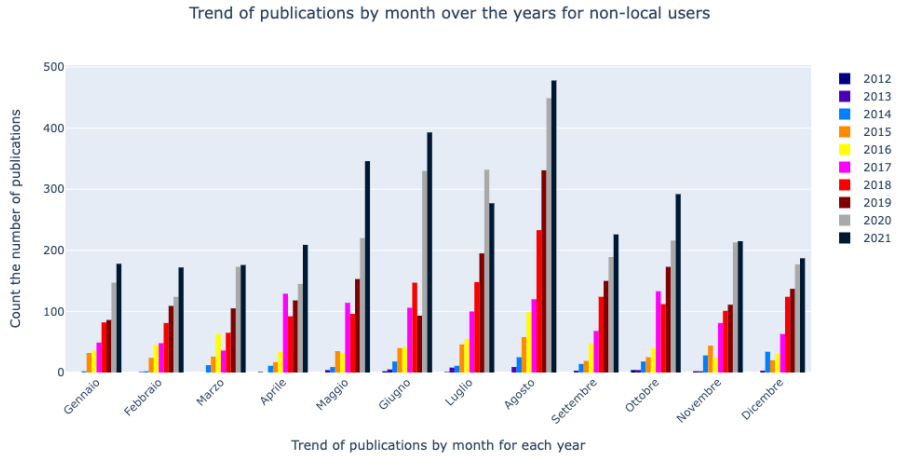

Figure 27. Trend of publications by month over the years for non-local users

Reprocessing the same data month by month in absolute values from 0 to 1 for all the years analyzed (Figure 28), there is a confirmation year by year of the seasonal trends which mainly concern the highest number of publications in the summer period. The considerations that can be obtained for the first three years analyzed (2012, 2013 and 2014) are of little importance, since the number of posts per year is very low and therefore it is easy that a single user having published more can significantly influence the general statistics of the year.

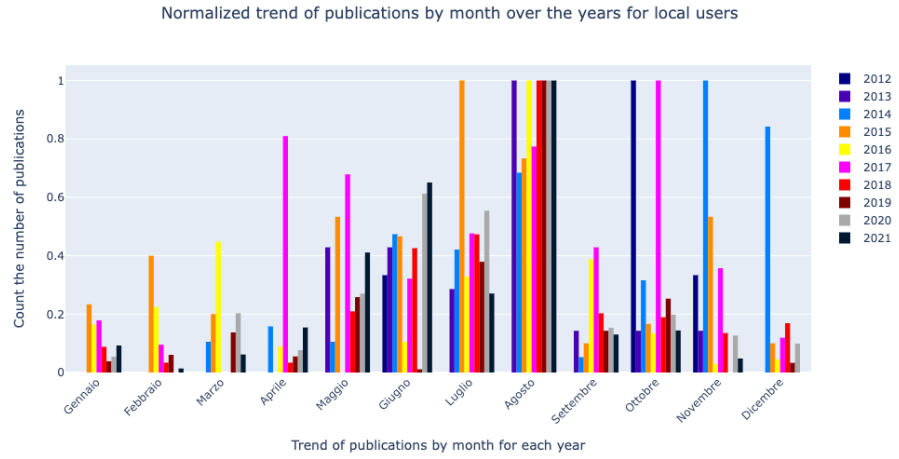

Figure 28. Normalized trend of publications by month over the years for local users

The above applies to both the graph in absolute values of local and non-local users (Figure 28 and 29). From the view of the graphs in absolute values, it is possible to see how important the impact of the summer period on publications is, in fact the difference is considerable between that time for the years that have above all a significant statistical value after 2017.

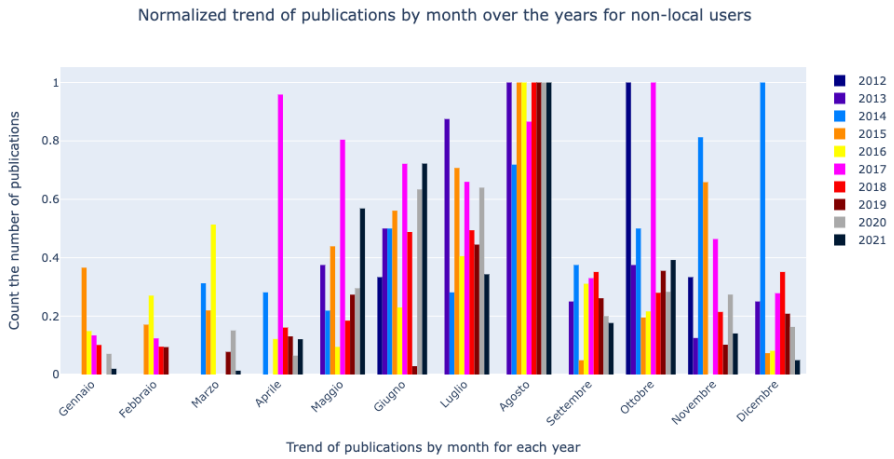

Figure 29. Normalized trend of publications by month over the years for non-local users

4.2.3 What places do locals and non-locals users frequent more seasonally?

From the data, graphs were drawn to understand if there was a difference in popularity in the places reported in the posts per season and between local and external people (Figure 30 and 31). First, in addition to what has already been seen on the total, a general division has been made between local and non-local and then seasonally.

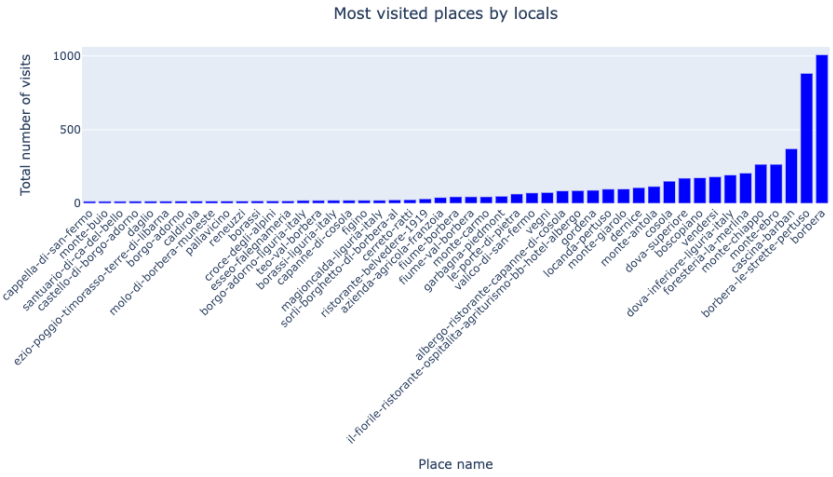

Figure 30. Most visited places by locals

There are differences in the places visited seasonally, that is between Spring, Summer, Autumn, and Winter. Even if a place is visited with a similar trend in different seasons, the reasons for the visit are for different reasons.

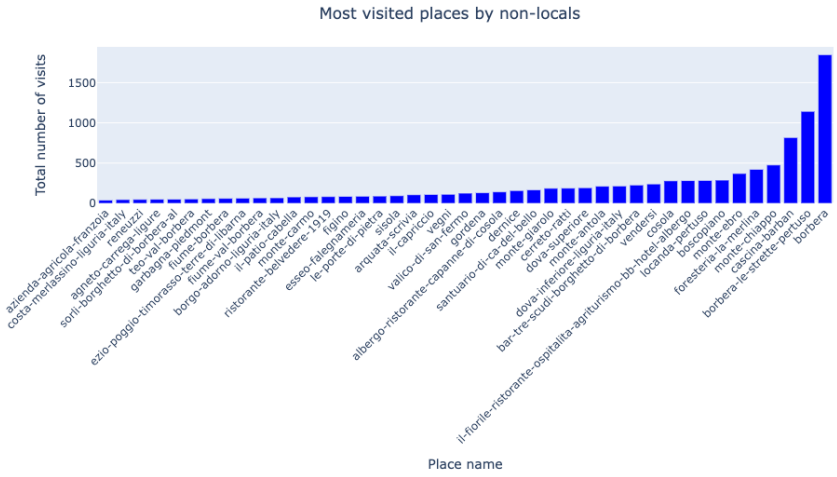

Figure 31. Most visited places by non-locals

From Figure 30 and Figure 31, the places most visited during the year by both locals and non-locals are generally the same for all seasons:

- the Borbera river and the Strette di Pertuso (river gorge about 6 kilometres long and even 100 meters high in some points, of considerable importance natural for the Piedmont Region), which are however reported as two different things even if the Strette are located inside the Borbera, but depending on how the user enters the hashtag it could be that in the post he referred to this natural area or another part because popular in the summer for swimming and spending the day.
- There are the Mountains, Monte Chiappo which is the top of the so-called Massiccio dell'Antola. This name is among the most mentioned together with another peak that is part of the same, Monte Ebro.
- Among the most reported restaurants in the geotags, there are Foresteria La Merlina in the municipality of Dernice, which is located not far from the Strette di Pertuso, above these there is also Locanda Pertuso. Finally, among the most mentioned there is also Il Fiorile located at the gates of the Lower Valley, in the hamlet of Torre Ratti in the municipality of Borghetto Borbera.
- Among the places of considerable impact, there is also the Cascina Barban which became famous thanks to its owner, a singer who managed to have some success nationally with the band Ex-Otago and says he found a point of inspiration in that place. Here he has a small farm with animals and a small restaurant. He organizes parties especially in the summer becoming a point of reference for the community, and furthermore, he is very committed to promoting the area.

- Another place mentioned in an important way is Boscopiano, it is located right above the Strette di Pertuso from which you can reach the river through a path. Often people gather there to have an aperitif or picnic in the summer.

We went to see if there was a difference in the visit of the places between locals and non-locals. The following shows the visits for the premises in the sequence of the seasons: spring, summer, autumn and winter.

Figure 32. Most visited places in spring by locals

In spring (Figure 32), there are no important differences between the most published places compared to what is seen in the general statistics. However, a locality “le-porte-di-pietra” stands out, which is not clearly detectable among the main places visited. By verifying what it is, it is not a geotag referring to a specific place, but it is a Mountain running competition held in the month of May (*Porte Di Pietra | Trail Running su lunghe distanze*, n.d.).

Even in summer (Figure 33), the data generally do not diverge much from the general ones, however, visits double for the two most visited places (Borbera and Strette di Pertuso). These places are an important destination for cooling off in the river water on weekends, which is why they also have much higher numbers this season than all other places.

A place generally ranked lower appears: dova-superiore, here there are two monuments dear to the local population, one is a very photographed church in summer especially during the night because it is very isolated and its lights contrast with the starry sky, while the other is a small chapel built for the partisans who fought against the fascists and Nazis during the end of the Second World War.

Dova Inferiore, part of the same municipality as the place mentioned above, is also particularly popular in this period, as several hiking trails watching the posts seem to start from here.

Figure 33. Most visited places in summer by locals

In autumn (Figure 34), the number of people who visit the Strette di Pertuso in particular, but also the Borbera are drastically reduced, in fact, due to the colder temperatures, people no longer go for a swim. In this period, indoor places such as restaurants and cascina-barban became more popular. As a result, for the first time this season Boscopiano reaches one of its lows.

Figure 34. Most visited places in fall by locals

In winter (Figure 35) there is the lowest figure compared to all the others, the Borbera river remains the most cited of all the geotags, but the Strette di Pertuso decrease in position for the first time among all the other seasons. In this season, there is a strong presence of people in the mountains, looking at the photos we discover the reason: people go to photograph the landscapes

whitened by snow. However, indoor places such as restaurants and cafes remain popular, in suburban places, they are often a meeting place, especially after work, before returning home.

Figure 35. Most visited places in winter by locals

For the spring (Figure 36), the first three places most visited by non-locals have generally double values compared to locals (the highest has 500 versus about 250), even if from the fourth onwards there are similar values. The first three places most visited by non-locals have generally double values compared to locals, even if from the fourth onwards there are similar values.

Figure 36. Most visited places in spring by non-locals

For spring, in addition to the natural spots, they remain popular among non-local dining venues and remain among the most visited places albeit in different locations. However, looking at the

locations with the highest number of visits, one can see a lower impact of this category on non-locals than on locals

The Borbera and the Strette di Pertuso have important numbers during the summer for non-locals (Figure 37). Restaurants also appear to be very popular; it could be one of the main reasons why non-locals visit the study area.

Figure 37. Most visited places in summer by non-locals

In autumn (Figure 38), closed places reach more important numbers. It appears that arquata-scrivia is a hashtag referring to a city outside the valley, but at its gates. The top three most popular places, especially the first one, have a very different trend than the other places. However, less for the same season for the same chart corresponding to the premises.

Figure 38. Most visited places in fall by non-locals

The peaks as regards the places most visited in winter by non-locals are much more pronounced, around 300 while for locals it remains around 100 (Figure 39). Winter is, therefore one of the seasons where there is the greatest difference, especially for other less visited places, which however remain with higher values as you go towards the bottom of the ranking.

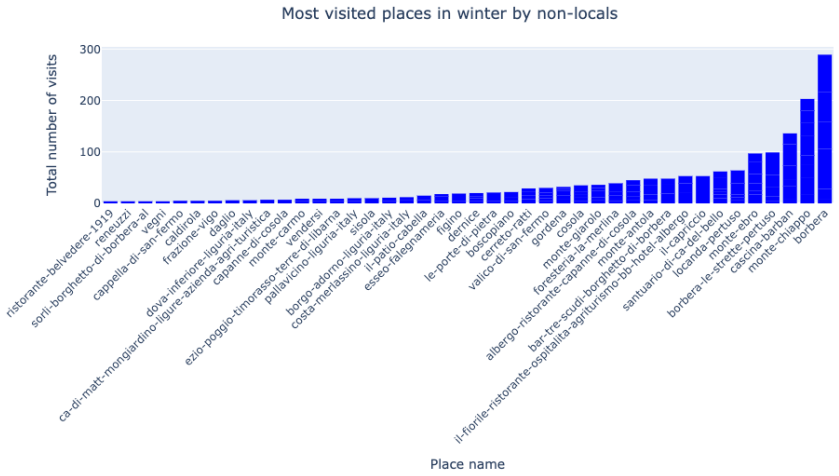

Figure 39. Most visited places in winter by non-locals

4.3 Are there clusters and hotspots?

From the construction of the tables, it was possible to create the diagrams we will see shortly to verify how the flows occur within the study area. The nodes have been georeferenced through the coordinates inserted as attributes to characterize them. It is possible to obtain various information. The color of belonging for the month in which it is most visited has been entered in the legend. The color that connects the edges refers to the target to which it is directed.

It is not possible through GEPHI to make an overlap with other georeferenced data, but it is possible to note that most of the points are concentrated along the valley floor. Few are the most visited places, and as we will see later the ones seen in the previous paragraphs remain. Despite this, some of the points, even if little visited, are concentrated in specific areas thus bringing the flow of people to concentrate in specific places, in this case above all in the valley floor of the more populous lower valley (Figure 40).

The graph was obtained through two algorithms made available by GEPHI, the first through a plugin called Geotag, allows you to georeferenced each node by inserting latitude and longitude, the second algorithms is called Noverlap, which distances the neighboring nodes to make them visible but losing the exact location.

About 50% of the movements of people between places that have a popularity especially in the summer period of which 20% in August. The statistics, however, are slightly different from what has already been seen previously as already mentioned because some elaborations have been made and there are only networks for users who have visited at least two places in the analyzed period. however, it is difficult to identify a cluster in this way because it seems that places visited in different periods are mixed, only for July we see a concentration of the part to the south-east.

Seeing the same GEPHI graph for the major activities carried out, we see a clear prevalence of nature, but for sport and food we see a concentration in specific places (Figure 41).

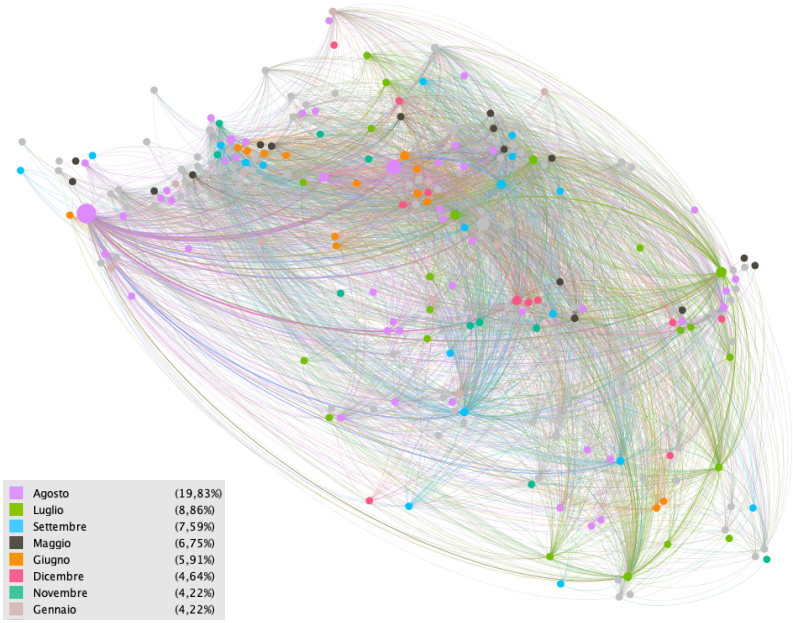

Figure 40. GEPHI diagram with point size for the number of views and color for the highest monthly visit

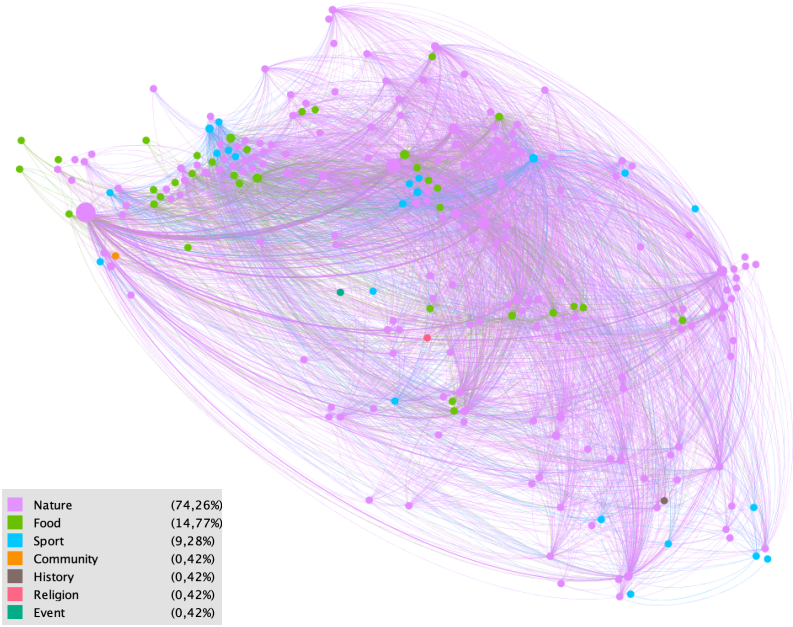

Figure 41. GEPHI diagram with point size for the number of views and color for the highest activity

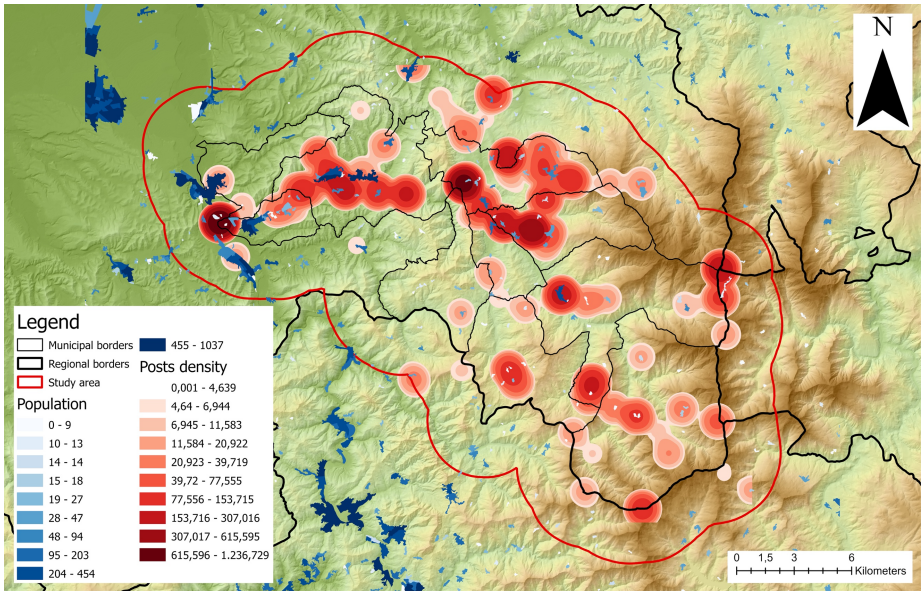

Figure 42. Location popularity overlaid on population

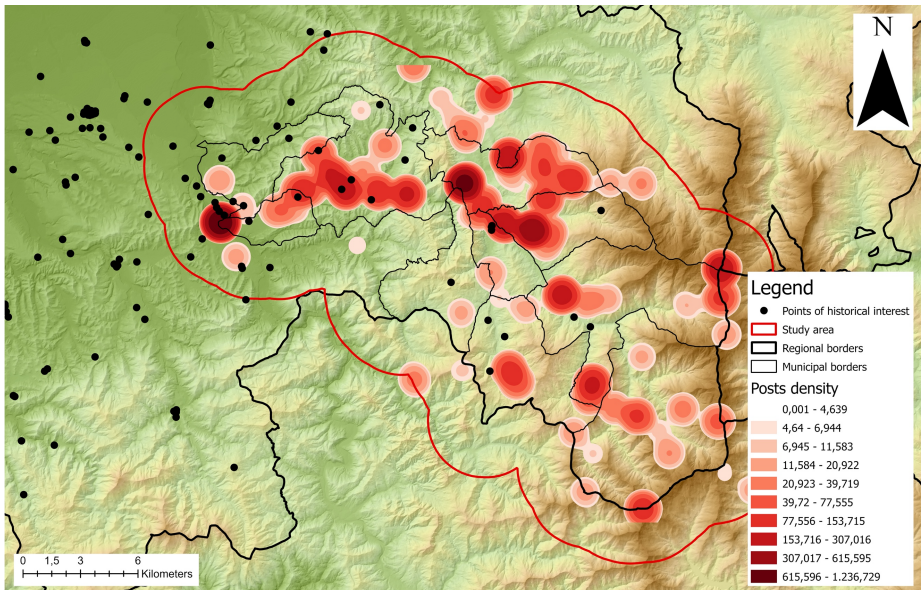

Figure 43. Post density and historical points of interest

From the map created through the ArcGIS Pro kernel density tool produced by entering the locations and the number of occurrences as a weight. It is possible to see how their impact is distributed (Figure 42). The highest density corresponds to the most populous areas, except for

some peripheral places such as in the mountainous areas to the south-east where the mountain peaks reported in the posts are found. Some geotags, however, are difficult to estimate, such as those referring to the Borbera river, which have an impact on the map in a specific place when the photos could have been taken for the entire perimeter of the river.

There seems to be little correlation when looking at the map on historical places with respect to the density of published posts (Figure 43).

The following image (Figure 44) is the result of a cleaning of the nodes that have an annual number of views lower than 47 which was the number seen by the statistics of the previous paragraphs in which the most significant places are found. This was done to give a clearer idea for at least the main locations how movements occur within the study area.

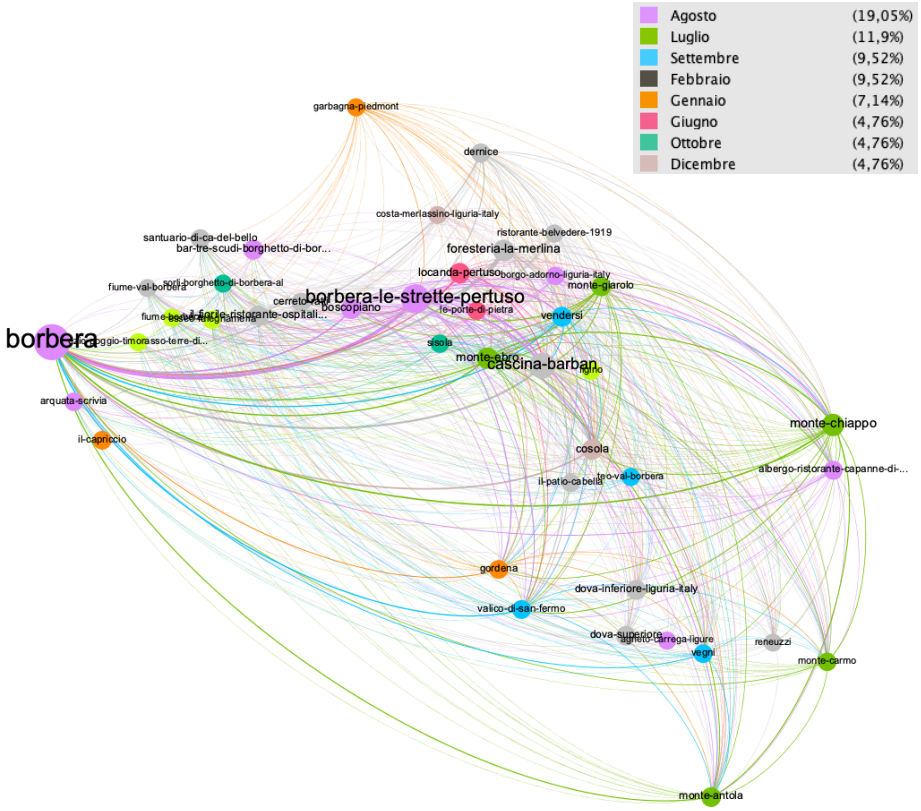

Figure 44. Nodes filtered for the highest number of visits and statistically more significant

5 Discussion

The purpose of this thesis was to apply scraping techniques no longer to a densely populated area, as learned from most of the research papers during the literature review phase, but to a peripheral area. The choice of the area wanted to be in continuation with the work done during the previous master thesis of the author for this reason, the choice fell on the Val Borbera.

One of the other objectives was to identify the social media that could best satisfy the needs regarding the information that was wanted to be recovered and the possibility of finding data. Among the various platforms found, the one that best suited the research wanted to do was Instagram, even if it was found that the amount of data that could be analyzed and downloaded was no longer possible as it had been for previous works.

However, these problems were resolvable since the research was done on a small and peripheral area, but despite this we arrived at a statistically significant sample with interesting trends, this applies to the aggregate data, but not for all the years of the obtained period. Furthermore, from the respect of the previously mentioned thesis, it was possible to focus the knowledge of at least a part of the territory.

The whole path for the construction of this thesis has always been uncertain and without a certainty of arriving at having data that could allow to have a minimum understanding of the territory. In particular the question concerned whether it was possible through a hashtag referring to a territory. Unfortunately, it was not possible in retrospect to know which data were found and the methodology as already mentioned is generally very poorly explained in academic papers. Furthermore, the information scraping part was particularly difficult trying not to waste time and make it as efficient as possible.

Despite the high number of academic papers found, it was very difficult to build a method that could obtain data from the platform due to the limitations, and few of the documents found gave an effective methodology. Some of them offered ready-made programs or cited some of those used, so, one of the solutions was to go inside each package and think about how to adapt each functionality to the needs. The achievement of a methodology took a long time and several attempts, it requires more experience even at the programming level than that of the author who had to obtain it.

One of the most difficult steps in the data analysis was the lack of differentiation in Instagram between different areal dimensions, for example between point and boundaries, and therefore it was necessary to build a model to report everything to punctual places. The model worked, but not optimally, without the limits imposed by the platform it would have been possible to do a much more accurate and thorough job. This is because sometimes the classification was not always achievable and for this reason one of the future implementations could be trying to build a model also on the classification of the images. Even if up to now it is not possible to carry out this action for Instagram and in any case privacy and law would prevail.

Building a model also based on photos would allow to use posts discarded for lack of a description. It could be particularly useful in urban areas or for items that are easily distinguishable from a model.

The usefulness of classifying the images from the social media could stop for improving the classification of the images and lead to further developments. An example could be the construction of urban landscape models to understand which element is prevalent within the city (green spaces, low visibility of the sky in the case of tall buildings, measure crowding in places, etc.) as it was done in a project, but using google street view (Middel et al., 2019). Another future development could be to investigate the meteorological phenomena through the observations of the people in the post descriptions.

A further step could be the implementation of more advanced techniques using hashtags. One of the examples would be trying to produce a model to determine the business, location, and other types of information that could be gleaned. The advantage would be since a hashtag has a precise meaning and for that reason the classification should be just as precise.

Another consequence of the limitations imposed by Instagram is the inability to obtain a large amount of data. This would be useful to deepen each place found to have a broader view on the trends and verify the validity of what was found only through the hashtag. It would probably have been possible to carry out this action, but it would have required several weeks of work, which is not the time available for the development of this thesis.

The hashtag #valbobera is interesting as it is gaining popularity recently, a further study step could be trying to understand the reason for this event. However, it is interesting to note how the trend has grown considerably during the covid years between 2019 and 2020 in that period restrictions were enacted. It is likely that many of the people visited places closer to their homes more frequently.

Other limitations encountered were the impossibility through the hashtag #valborbera to find all the locations reported by users on Instagram because the analyzed hashtag is not always used when the publication is made, but only the location is entered. It is also interesting to note that the number of external users is much higher than the local ones or those coming from around the area, but how much the publications are higher for local users. However, it is very difficult to have a general vision through the network analysis of the movements made by non-locals because most of them have only carried out a movement within the area and therefore it is necessary to think about other types of visualization of their movements. The data suggest a significant attendance of people not directly from the place, but mostly from surrounding areas from the surrounding provinces, the one in which the area is located and others also coming from more distant. Different influxes to places have been found according to the seasons, but generally the most visited places remain so throughout the year, only the frequency of visits changes.

Another point found to investigate that would have required to go location by location is to verify the target of people who use the hashtag #valbobera. In fact, it seems that its use, is mainly done by people interested in outdoor activities. This is also confirmed by the fact of the latest considerations made through the network analysis made. The places where people have gone there for an interest in nature decrease in the winter season, and then return in the summer. Some places remain popular especially to capture the characteristics of the landscape during periods and meteorological events such as snow.

Generally, the area is most frequented during the summer season even if analyzing specifically the activities most carried out by people which are sport, food and nature, nature being the most

numerous categories greatly influences the rest of the sample, while sport it has a relatively constant trend throughout the year with a peak around May, while food has two peaks in May and October.

At the same time, the movements are concentrated around a few main places that contain most of the flow of visitors in the area. It is interesting to note from the eyes of a person who knows the place that some of the important transit places of people within the area are not very marked. Interesting is the case of a pizzeria located in Vignole Borbera where there has been a strong influx of people for decades now, however, as the owner is not active on social media, it is not advertised and therefore cannot be found on Instagram if not looking for the Geotag that anyway has a considerable number of references. A possible solution to this type of problem could be to investigate more deeply the places visited at least by the locals to try to identify the greatest number of places. Analyzing the places year by year allows one to understand and perhaps detect some dynamics such as the end or the beginning of activities, the beginning, and the end of popular events, etc. Some geotags, however, are difficult to estimate, such as those referring to the Borbera river, which have an impact on the map in a specific place when the photos could have been taken for the entire perimeter of the river. There are other very important places, one is a Water Park among the largest in northern Italy and very busy especially by people outside the area even outside the surrounding provinces. Another undetected place of considerable impact on the territory is the Il Mulino nightclub, frequented in large numbers by young people who certainly do not care to indicate the context in which it is inserted.

From the experience carried out it emerged that it is important to analyze the data at least seasonally, see the example of the “le-porte-di-pietra” locality which proved not to be a place, but an event that is held every year. To identify other events of this type it would be interesting to analyze these data even more specifically, perhaps by month.

6 Conclusion

Starting from the search for a methodology, we tried to adapt works already made to the purpose of this thesis developing a code and its improvement to make it more efficient and perfectly adaptable to the objectives to achieve. As a result, it was possible to get some interesting feedbacks.

The bibliographies currently available are many, but prior to the limitations imposed by the company that bought Instagram, new methods of analysis must be discovered, and it is no longer possible to use the same amount of data as before. Therefore, in this sector, the development of social media analysis has slowed down significantly.

Despite this, a process was built, it took a lot of time, but it was not possible to detect all the dynamics that take place within the study area. Therefore, we have got only a general vision and linked to a specific interest of the people who have published. It was found that the analyzed hashtag is strongly targeted by people whose interest is to carry out outdoor activities, both by local and non-local people. Most external users have published only one post and they are many more than the local who, however, publish a lot. Moreover, it has been found that there are light seasonal differences between these two groups and different interests in the activities.

The most popular season is summer, and the movement of people takes place in specific places in the Borbera valley, some of which are the main. This has been possible to make visible with the diagrams constructed by GEPHI, a program useful for network analysis.

The problem with applying analysis like made in this work is that we do not have a sample that is statistically consistent, as we have seen, it has happened for some years. Being in a low density and not highly popular area, there has been a high targeting of people's interest in nature activities. We could not find another popular platform in the same way as Instagram and with the ability to extract the same type of information. For this reason, it was not possible to compare the information obtained.

The accuracy of the classification could have been improved by comparing additional information such as using the photos.

7 References

Abdulrahman, R., Neagu, D., Holton, D. R. W., Ridley, M., & Lan, Y. (2013). Data extraction from online social networks using application programming interface in a multi agent system approach. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8065*. doi: 10.1007/978-3-642-41776-4_4

Aimar, A. (2020). *INSTA-TURIN Revising the notion of space, community and proximity in the contemporary city through Instagram social urban data*.

Akhtar, N. (2014). Social network analysis tools. *Proceedings - 2014 4th International Conference on Communication Systems and Network Technologies, CSNT 2014*, 388–392. doi: 10.1109/CSNT.2014.83

AlSayyad, N., & Guvenc, M. (2015). Virtual Uprisings: On the Interaction of New Social Media, Traditional Media Coverage and Urban Space during the 'Arab Spring.' *Urban Studies*, *52*(11). doi: 10.1177/0042098013505881

Amatori, F. (2013). *L'IRI dagli anni Trenta agli anni Settanta in "Il Contributo italiano alla storia del Pensiero: Tecnica."* Enciclopedia Treccani. Retrieved from https://www.treccani.it/enciclopedia/l-iri-dagli-anni-trenta-agli-anni-settanta_%28Il-Contributo-italiano-alla-storia-del-Pensiero:-Tecnica%29/

Babar, M., & Arif, F. (2017). Smart urban planning using Big Data analytics to contend with the interoperability in Internet of Things. *Future Generation Computer Systems*, *77*. doi: 10.1016/j.future.2017.07.029

Barkhuus, L., & Tashiro, J. (2010). Student socialization in the age of facebook. *Conference on Human Factors in Computing Systems - Proceedings*, *1*. doi: 10.1145/1753326.1753347

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third International AAAI Conference on Weblogs and Social Media*. doi: 10.1136/qshc.2004.010033

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, *28*. doi: 10.1016/j.inffus.2015.08.005

Boy, J. D., & Uitermark, J. (2016). How to study the city on instagram. *PLoS ONE*, *11*(6). doi: 10.1371/journal.pone.0158161

Boy, J. D., & Uitermark, J. (2017). Reassembling the city through Instagram. *Transactions of the Institute of British Geographers*, *42*(4). doi: 10.1111/tran.12185

Camacho, D., Panizo-LLedot, Á., Bello-Orgaz, G., Gonzalez-Pardo, A., & Cambria, E. (2020). The four dimensions of social network analysis: An overview of research methods, applications, and software tools. *Information Fusion*, *63*. doi: 10.1016/j.inffus.2020.05.009

Candia, C., Encarnação, S., & Pinheiro, F. L. (2019). The higher education space: connecting degree programs from individuals' choices. *EPJ Data Science*, *8*(1). doi: 10.1140/epjds/s13688-019-0218-4

Chen, Y., Parkins, J. R., & Sherren, K. (2018). Using geo-tagged Instagram posts to reveal landscape values around current and proposed hydroelectric dams and their reservoirs. *Landscape and Urban Planning*, *170*. doi: 10.1016/j.landurbplan.2017.07.004

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). The Livehoods project: Utilizing social media to understand the dynamics of a city. *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*.

Cunietti, S. (2019). *Patrimonio culturale, paesaggio e pianificazione nell'Alto Monferrato*. Retrieved from https://webthesis.biblio.polito.it/10138/

Domènech, A., Mohino, I., & Moya-Gómez, B. (2020). Using flickr geotagged photos to estimate visitor trajectories in world heritage cities. *ISPRS International Journal of Geo-Information*, *9*(11). doi: 10.3390/ijgi9110646

Domínguez, D. R., Díaz Redondo, R. P., Vilas, A. F., & Khalifa, M. ben. (2017). Sensing the city with Instagram: Clustering geolocated data for outlier detection. *Expert Systems with Applications*, *78*. doi: 10.1016/j.eswa.2017.02.018

Felice, E. (2015). *Lo sviluppo economico delle regioni: dalle tre Italie alle due Italie in "L'Italia e le sue Regioni."* Enciclopedia Treccani. Retrieved from https://www.treccani.it/enciclopedia/lo-sviluppo-economico-delle-regioni-dalle-tre-italie-alle-due-italie_%28L%27Italia-e-le-sue-Regioni%29/

Gosal, A. S., Geijzendorffer, I. R., Václavík, T., Poulin, B., & Ziv, G. (2019). Using social media, machine learning and natural language processing to map multiple recreational beneficiaries. *Ecosystem Services*, *38*. doi: 10.1016/j.ecoser.2019.100958

Grandjean, M. (2015). *GEPHI – Introduction to Network Analysis and Visualization*. Retrieved from http://www.martingrandjean.ch/gephi-introduction/

Guerrero, P., Møller, M. S., Olafsson, A. S., & Snizek, B. (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). doi: 10.17645/up.v1i2.609

Hristova, D., Williams, M. J., Musolesi, M., Panzarasa, P., & Mascolo, C. (2016). Measuring urban social diversity using interconnected geo-social networks. *25th International World Wide Web Conference, WWW 2016*. doi: 10.1145/2872427.2883065

Hudson, S., Roth, M. S., Madden, T. J., & Hudson, R. (2015). The effects of social media on emotions, brand relationship quality, and word of mouth: An empirical study of music festival attendees. *Tourism Management*, *47*. doi: 10.1016/j.tourman.2014.09.001

Ibáñez-Cubillas, P., Díaz-Martín, C., & Pérez-Torregrosa, A. B. (2017). Social Networks and Childhood. New Agents of Socialization. *Procedia - Social and Behavioral Sciences*, *237*. doi: 10.1016/j.sbspro.2017.02.026

Instagram - Wikipedia. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Instagram

Instagram Platform. (n.d.). Retrieved from https://developers.facebook.com/docs/instagram?locale=en_US

Itani, O. S., Agnihotri, R., & Dingus, R. (2017). Social media use in B2b sales and its impact on competitive intelligence collection and adaptive selling: Examining the role of learning orientation as an enabler. *Industrial Marketing Management*, *66*. doi: 10.1016/j.indmarman.2017.06.012

Kim, A. J., & Ko, E. (2012). Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. *Journal of Business Research*, *65*(10). doi: 10.1016/j.jbusres.2011.10.014

Latorre-Martínez, M. P., Iñíguez-Berrozpe, T., & Plumed-Lasarte, M. (2014). Image-focused social media for a market analysis of tourism consumption. *International Journal of Technology Management*, *64*(1). doi: 10.1504/IJTM.2014.059234

Lee, J. G., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, *2*(2). doi: 10.1016/j.bdr.2015.01.003

Legge 23 ottobre 1859 n. 3702 - Wikipedia. (n.d.). Retrieved from https://it.wikipedia.org/wiki/Legge_23_ottobre_1859_n._3702

Mamei, M., Rosi, A., & Zambonelli, F. (2010). Automatic analysis of geotagged photos for intelligent tourist services. *Proceedings - 2010 6th International Conference on Intelligent Environments, IE 2010*. doi: 10.1109/IE.2010.34

Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2019). Social Media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, *74*. doi: 10.1016/j.compenvurbsys.2018.11.001

Middel, A., Lukasczyk, J., Zakrzewski, S., Arnold, M., & Maciejewski, R. (2019). Urban form and composition of street canyons: A human-centric big data and deep learning approach. *Landscape and Urban Planning*, *183*. doi: 10.1016/j.landurbplan.2018.12.001

Molinari, L. (1999). *Piano, porto, città. L'esperienza di Genova*. Autorità Portuale Di Genova/Skira. Retrieved from https://www.archiviogabrielebasilico.it/it/scopri/lavori/piano-porto-citta-lesperienza-di-genova

Oliveira, E., & Panyik, E. (2015). Content, context and co-creation: Digital challenges in destination branding with references to Portugal as a tourist destination. *Journal of Vacation Marketing*, *21*(1). doi: 10.1177/1356766714544235

Porte Di Pietra | Trail Running su lunghe distanze. (n.d.). Retrieved from https://www.portedipietra.it/

Rathore, M. M., Ahmad, A., Paul, A., & Rho, S. (2016). Urban planning and building smart cities based on the Internet of Things using Big Data analytics. *Computer Networks*, *101*. doi: 10.1016/j.comnet.2015.12.023

Rodriguez, M., Peterson, R. M., & Krishnan, V. (2012). Social media's influence on business-to-business sales performance. In Journal of Personal Selling and Sales Management (Vol. 32, Issue 3). doi: 10.2753/PSS0885-3134320306

Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame.

Proceedings of the National Academy of Sciences of the United States of America, *111*(52). doi: 10.1073/pnas.1410931111

Wang, X., Yu, C., & Wei, Y. (2012). Social Media Peer Communication and Impacts on Purchase Intentions: A Consumer Socialization Framework. *Journal of Interactive Marketing*, *26*(4). doi: 10.1016/j.intmar.2011.11.004

Yu, C. E., & Sun, R. (2019). The role of Instagram in the UNESCO's creative city of gastronomy: A case study of Macau. *Tourism Management*, *75*. doi: 10.1016/j.tourman.2019.05.011

Zhai, S., Xu, X., Yang, L., Zhou, M., Zhang, L., & Qiu, B. (2015). Mapping the popularity of urban restaurants using social media data. *Applied Geography*, *63*. doi: 10.1016/j.apgeog.2015.06.006

Masters Program in **Geospatial Technologies**

Supported by:
Education and Culture
ERASMUS MUNDUS