

Masterarbeit

Extraction of places of interest from VGI

eingereicht von **Junru Lin**
geboren am 10.10.1996 in Liaoning, China

zum Erlangen des Hochschulgrades
Master of Science (M.Sc.)

Tag der Einreichung 05.11.2021

Betreuer Dr.-Ing. Alexander Dunkel
Technische Universität Dresden

Mathias Gröbe M.Sc.
Technische Universität Dresden

Original Task Description

The task for this thesis is to explore two nationwide spatial datasets collected from Flickr and Instagram (including Facebook Places) to extract collective places of interest, to be used in a web map (tile layer). The candidate will further explore how typical metrics used in VGI can be used to classify places regarding their collective value of interest. Depending on the information content available, the two datasets may be combined with additional context data collected from other sources such as Geonames, OpenStreetMap, or Wikidata.

Statement of Authorship

Herewith I declare that I am the sole author of the thesis named

"Extraction of places of interest from VGI"

which has been submitted to the thesis assessment board today.

I have fully referenced the ideas and work of others, whether published or unpublished.
Literal or analogous citations are clearly marked as such.

Dresden, 05/11/2021

Signature



Cartography M.Sc.

Master thesis

Extraction of places of interest from VGI

Junru Lin



2021

Extraction of places of interest from VGI

submitted for the academic degree of Master of Science (M.Sc.)
conducted at the Institute of Cartography, TU Dresden

Author:	Junru Lin
Study course:	Cartography M.Sc.
Supervisor:	Dr.-Ing. Alexander Dunkel (TUD) Mathias Gröbe M.Sc. (TUD)
Reviewer:	Francisco Porras Bernárdez M.Sc. (TUW)

Chair of the Thesis
Assessment Board: Prof. Dipl.-Phys. Dr.-Ing. habil. Dirk Burghardt

Date of submission: 05.11.2021

Statement of Authorship

Herewith I declare that I am the sole author of the submitted Master's thesis entitled:

"Extraction of places of interest from VGI"

I have fully referenced the ideas and work of others, whether published or unpublished. Literal or analogous citations are clearly marked as such.

Dresden, 05.11.2021

Junru Lin

Acknowledgments

First of all, I would like to express my most sincere gratitude to my supervisors: Dr.-Ing. Alexander Dunkel and Mathias Gröbe M.Sc.. They have given me inspiring guidance, helpful advice, instant feedback, endless patience, and encouragement during this period.

Also, thank you to my whole thesis assessment board, who provided me with valuable advice and feedback in the proposal defense and the midterm presentation. Thank you to all the staff of Cartography M.Sc., especially the coordinators at the four universities, whose work and attentive help have been essential for these more than two years of study and life. I am genuinely grateful for your efforts and dedication, especially during the challenging epidemic situation.

Thank you to my classmates since many of my memories with you have been an integral part of my graduate studies. Thank you to all my friends in Germany and China who have given me more support and comfort than you can imagine. Thank you to my boyfriend Yves, who took good care of me when I was ill and helped me recover physically and mentally.

Finally, thanks to my parents. Without your help, coming to Europe to study would not have been possible. We have not seen each other for more than two years because of the epidemic, but your love from more than 7,000 km distance and 7 hours' time difference has never been absent.

Abstract

In many application scenarios, such as urban planning, traffic guidance, travel planning, POI plays a vital role in supporting decision making. With almost everyone using social media and the widespread use of big data, vast amounts of data from social media are increasingly being analyzed and visualized for a variety of purposes, such as marketing, education, and polling. Following this trend, traditional questionnaires, fieldwork, and other methods to count POI information are gradually replaced by VGI data from OSM and social media, including Flickr, Facebook, Instagram, etc. This provides new possibilities for POI extraction and description methods.

Therefore, this research aims to develop a workflow to visualize and summarize POIs or AOIs for travel planning purposes, on a multi-scale and national-range interactive web map, based on three VGI datasets within Germany. In order to achieve that, this study combines three spatial data aggregation methods: grid-based aggregation, administrative boundaries-based aggregation, and DBSCAN. The aggregated data are visualized on an interactive web map application supported by Mapbox GL JS API. This application contains heat maps, choropleth maps, and proportional symbol maps and uses different metrics, including Post Count, User Count, Post Count per capita. The known popular areas and tourist attractions can be identified with this method. The output web map allows the tourists and photographers to compare and explore the AOIs and POIs within Germany during their trip planning stage.

Keywords: VGI, POI extraction, LBSM, spatial data aggregation, interactive web map

Contents

Original Task Description	2
Statement of Authorship	3
Acknowledgments.....	7
Abstract.....	8
Contents.....	1
Figures.....	3
Tables.....	4
Abbreviations.....	5
1 Introduction.....	6
1.1 Motivation	6
1.2 Research Objective and Questions	7
1.3 Thesis Structure	8
1.4 Study Area	8
2 Theoretical Background	11
2.1 Places of Interest	11
2.2 Data Sources.....	12
2.2.1 Volunteered Geographic Information	12
2.2.2 Location-based Social Media.....	13
2.2.3 Privacy Issues	15
2.3 Data Processing and Visualization	16
2.3.1 HyperLogLog.....	16
2.3.2 Density-based Clustering.....	17
2.3.3 Interactive Web Map Application.....	19
3 Application.....	20
3.1 LBSN Data	21
3.1.1 Aggregated place name dataset	21
3.1.2 Aggregated post dataset	23

3.1.3	Flickr CCBY post dataset.....	24
3.2	Administrative data.....	26
3.3	Data Aggregation	27
3.3.1	Grid-based Aggregation	27
3.3.2	Administrative Boundaries-based Aggregation	30
3.3.3	Density-based Data Clustering	31
3.3.4	Subjective Information Processing.....	33
4	Interactive Visualisation	36
4.1	User Interface.....	36
4.2	Visualization for Overview.....	39
4.2.1	AOI Heatmap.....	39
4.2.2	AOI Choropleth Map.....	42
4.3	Visualisation for Local POIs	46
4.3.1	Study case in Dresden.....	46
4.3.2	Study case in Berlin	48
5	Discussion	51
5.1	Answers to the Research Questions.....	51
5.2	Evaluation of the privacy protection	53
5.3	Evaluation of Data Aggregation and POI extraction	54
5.4	Limitations of the Web Map.....	55
6	Conclusion and Outlook	57
6.1	Conclusion	57
6.2	Future work	57
	References.....	59
	Appendices.....	64
	Appendix A: SQL code for grid-based aggregation	64
	Appendix B: SQL code for administrative boundaries-based aggregation.....	65
	Appendix C: SQL code for tags ranking and filtering	66

Figures

Figure 1 Map of Germany (WorldAtlas, 2021).....	9
Figure 2 Create a new geofence on Flickr (Flickr blog, 2011).....	15
Figure 3 An example of HLL shard structure	17
Figure 4 Illustration of the DBSCAN cluster model (Schubert et al., 2017).....	18
Figure 5 Workflow Diagram.....	20
Figure 6 Original aggregated post data within Berlin	28
Figure 7 Data within Berlin after aggregation.....	29
Figure 8 Data within Germany after NUTS boundaries-based aggregation	31
Figure 9 The user interface of LoFo including a navigation bar and a map display area ..	36
Figure 10 Diagram of how the search box works.....	37
Figure 11 The user interface of LoFo on a mobile device	38
Figure 12 Heatmap displaying the region around Saxony (at zoom level 8).....	40
Figure 13 AOIs around Füssen and Garmisch-Partenkirchen using PC metric.....	41
Figure 14 AOIs around Füssen and Garmisch-Partenkirchen using UC metric	41
Figure 15 The initial state of the choropleth map showing a summary of Germany.....	42
Figure 16 Choropleth map with an aggregation on NUTS1 level.....	43
Figure 17 Choropleth map with an aggregation on NUTS2 level	44
Figure 18 Choropleth map with an aggregation on NUTS3 level	45
Figure 19 Choropleth map with an aggregation on postcode.....	46
Figure 20 Local POIs in Dresden	47
Figure 21 Local POIs in Dresden city center.....	48
Figure 22 Local POIs in Berlin.....	49
Figure 23 Local POIs around Brandenburger Tor	50
Figure 24 Tourist attractions in Dresden Innere Altstadt	55
Figure 25 Map legends in choropleth maps	56

Tables

Table 1 Six sub-categories of social media (Kaplan & Haenlein, 2010)	13
Table 2 Comparison of four social networking applications.....	14
Table 3 Performance comparison for raw and HLL data processing (Dunkel et al., 2020)	16
Table 4 Description of three VGI datasets	21
Table 5 Data fields in aggregated place name dataset	22
Table 6 Comparison of data from three social networking applications.....	22
Table 7 Description of data fields in aggregated post dataset	24
Table 8 Description of data fields in Flickr CCBY post dataset.....	25
Table 9 Geohash length and distance of adjacent cell in meters (VGIscience, n.d.)	28
Table 10 Comparison of results after density-based clustering.....	32
Table 11 Zoom levels (Mapbox, 2021)	39

Abbreviations

AOI	Area of Interest
CSV	Comma-separated Values
DBSCAN	Density Based Spatial Clustering of Applications with Noise
GIS	Geographic Information System
GPS	Global Positioning System
HLL	HyperLogLog
IDBSCAN	Integrated DBSCAN
KNNDSCAN	K-nearest neighbors DBSCAN
LBSM	Location Based Social Media
LBSN	Location Based Social Network
MAU	Monthly Active Users
NUTS	Nomenclature of Territorial Units for Statistics
OSM	OpenStreetMap
PC	Post Count
POI	Place of Interest
PSBR	Point-set-based Region
PUD	User Days
SRID	Spatial Reference Identifier
UC	User Count
VDBSCAN	Varied DBSCAN
VGI	Volunteered Geographic Information

1 Introduction

1.1 Motivation

Places or Points of Interest (POIs) and Areas of Interest (AOIs, also called regions of interest [ROIs]) are essential information bases in many areas of decision making, such as for routing and urban planning purposes. It is important to note that POIs (and AOIs) typically combine two aspects of information. The first part consists of physically quantifiable properties of the environment, such as the location or thematic attributes of a POI. On the contrary, the second part is based on subjective, which is often hard to evaluate qualities of the environment (e.g., the popularity of places, measured through, e.g., visitation rates). By combining both of these aspects, POI- and AOI-based maps are beneficial for an extensive list of applications for certain groups of people.

Following the rapid development and pervasive use of location-based technology, traditional questionnaire-based approaches to collect and depict POIs such as field surveys or travel diaries can now be combined with or substituted by large volumes of spatio-temporal data. This offers new opportunities to visualize and understand urban dynamics and human movement (Arribas-Bel, 2014). For instance, GPS trace data such as taxi trajectory data from GPS-enabled taxis have been applied to define POIs and enhance POI information in various academic studies (Lerin et al., 2011; Yang & Ai, 2018; Liu et al., 2021). Other VGI (Volunteered Geographic Information) from various public platforms have also been utilized to explore POIs in plenty of researches (Popescu & Shabou, 2013; Yang et al., 2014; Corradi et al., 2015; Meng et al., 2017; Dunkel et al., 2019).

In this work, the improvement of POI- and AOI-based maps for tourists is specifically considered.

In this regard, two gaps in POI maps are of specific interest. Firstly, despite current trends in information visualization, POIs are still often displayed as pins on maps or as ranked lists of places for cities. However, there exist few map services that help tourists to find generally interesting AOIs, or to get small-scale overviews and summaries for areas, for example, during initial explorative trip planning. Secondly, popularity assessment is often ambiguous based on data sources that are limitedly representative. By including publicly available data sources, such as Volunteered Geographic Information (VGI) and Location Based Social Media (LBSM), the representativeness can be significantly improved.

Characteristics of POI information on public VGI or social media platforms arouse unsolved and challenging questions during POI data analysis and visualization. POIs are more abundant on social media platforms such as Facebook and Instagram than in mapping platforms like Google and OSM, which have strict quality control (Hochmair et

al., 2018). Social media platforms indeed contain a high volume of information that can help us to identify POIs and complement related visitor experiences. But compared with mapping platforms, a way to deal with the relatively ambiguous geographic information from LBSM should be figured out. For example, on Flickr, the accuracy of the geo-tagged information heavily depends on the number of items available in the corresponding area, which means that photos taken at highly popular areas can be more accurate (Hauff, 2013). However, this effect is mainly noticeable on the largest scale. The influences from this can be reduced by spatial aggregation when we zoom out to smaller scales. And lastly, privacy and sensitivity towards the use of data are an increasing concern, primarily when we work on LBSM data (Dunkel et al. 2020). Geo-tagged information from those social media platforms must be processed before we utilize it on any other applications to prevent user information leakage.

Based on this set of bounding questions, the main context of this research is hence developing a workflow to better use the abovementioned characteristics of VGI (notably LBSM data) for summarizing and aggregating POIs or AOIs on a multi-scale web map application. This research is based on three data sets collected from one of Instagram, Flickr, Facebook, and Twitter or a combination of several of them, from 2010 to 2020, covering Germany. Two data sets have been abstracted in a reduced, statistic data format that is particularly suited to visualize quantities and aggregation and that particularly prevents any identification of individuals. Questions explored herein specifically focus on smaller scales and summaries of data to reduce the influence of lower location accuracy and semantic deviation on the visualization results.

1.2 Research Objective and Questions

This research aims to develop a workflow to visualize and summarize POIs or AOIs for tourists, on a multi-scale and national-range map, based on data derived from VGI (notably LBSM data). This work attempts to achieve the following three research sub-objectives, and their corresponding specific questions (a) to (h) should be discussed and solved during the research process.

I . Identify the needs of visualizing POIs or AOIs for tourists and describe the data:

- (a) For what purposes are tourists using visualizations of POI or AOI, and what are the requirements on different map scales?
- (b) What are the pros and cons of combining data from multiple social media platforms for multi-scale extraction and visualization of POIs for tourists?
- (c) How is the data structured, and what is the volume of available data?
- (d) What parts of the data are related to either objective or subjective information?

II. Select approaches of summarizing and aggregating POIs or AOIs from VGI:

(e) What is the difference between different metrics, e.g., User Count, Post Count, User Days?

(f) What methods or algorithms should be employed while summarizing POIs or AOIs for different map scales?

III. Create the interactive visualization for the POIs and AOIs:

(g) How can POIs and AOIs be visualized on maps on different scales? Which information is important on which scale?

(h) Are there necessary map elements and map interactive actions that can be included while visualizing POIs for tourism purposes, and how will they need to be implemented in real scenarios?

1.3 Thesis Structure

This thesis is composed of six sections. The first section is an introduction section, which introduces the topic, states the motivation, and clarifies research objectives and questions. The structure of this thesis and an overview of the study area are also covered in this section. The second section lists and explains the background theories and methods that support this study. Topics, including places of interest, volunteered geographic information, theories regarding data processing and visualization such as HyperLogLog, typicality, density-based clustering, and web map, are investigated and examined. Following the background context, the third section is about applications. It describes three datasets that are utilized in this work, briefly explains the process of data preprocessing, and demonstrates the workflow of data aggregation for both small scale maps and large scale maps, in which Dresden is used as an example. After the data processing part, the interactive data visualization for both AOI-overview and local POIs is reported in the fourth section. In the fifth section, web maps, the output of the experiment, are evaluated. Additionally, the limitations of data aggregation and data visualization are discussed. The final section summarises this work and analyses what can be achieved in future research to lead the study to a larger context.

1.4 Study Area

The study area of this work is Germany, officially the Federal Republic of Germany. Germany is a country situated in central Europe, with a latitude range of 47°17' N to 55°03' N and a longitude range of 5°53' E to 15°2' E.

Topographically, Germany has incredible variety, as depicted in figure 1. Part of the Alps is located close to the southern border of the federal state of Bavaria, which makes this area an attraction for hikers throughout the year and a hotspot during skiing seasons. Besides Bavarian Alps, other forested hills such as Black Forest, Thuringian Forest, the Bohemian Forest, and mountains like the Ore Mountains and the Vogelsberg Mountains distribute across the central and southern regions of Germany. These areas are also quite popular as vacation destinations, which provide people peace away from fast-paced city life. In northern Germany, the landscape flattens as a broad plain extending to the North Sea. For instance, within this part, Spreewald, consisting of forested areas and wetlands crossed by canals, provides excellent chances for people to paddle and get close to nature.



Figure 1 Map of Germany (WorldAtlas, 2021)

Germany is composed of 16 federal states, which share a common culture, but each has its own characteristics. Fertilized by various cultures and rich pasts, numerous cities and towns are as well worth visiting and attract tourists regardless of the season. Cities

such as Berlin, Munich, Cologne, and Dresden have always been listed in travel recommendations when tourists would like to experience the culture and history in Germany.

Having so many possibilities, Germany attracts millions of tourists every year. According to a survey conducted in 2020, only approximately 13 percent of the holidaymakers did not book anything in advance (Statista, 2021). Other 87 percent tend to book accommodation, holiday packages, and/or tickets before departure. Information regarding the popularity of places of interest, accommodation prices, etc., is needed during different stages of planning.

2 Theoretical Background

2.1 Places of Interest

Before digging into the definitions of points of interest, places of interest, and areas of interest, which will be discussed in this work later, three nouns should be distinguished: point, place, and area. This work ranks these three terms in order of spatial scale from smallest to largest as point, place, and area. Generally, an area delivers a sense of a region on a surface, such as a city, a town, or a district, while a place is somewhere within a region (an area). Examples of places include a children's playground, a shopping mall, a Christmas market, and so on. Compared with area and place, a point has an even smaller spatial scale which refers to a precise location (sometimes with a specific latitude and longitude on the earth). For instance, a bench in a park, a statue, and a hotel can be regarded as points in this work. Although there is no strict differentiation between these three nouns, such as in kilometers or meters, in the context of this research, it is reasonably clear to discriminate them.

Instead of areas of interest (AOIs), urban areas of interest, which are areas within the urban environment having a greater degree of public engagement, are mentioned more frequently in past studies (Hu et al., 2015; Liu et al., 2021). In cities, because of the high population density and limited resources, urban AOIs are usually given higher priorities and more attention in urban planning projects and traffic analysis. Additionally, since urban AOIs present personal or public interests, they also play an essential role in personal or general tour planning and travel recommendations (Liu et al., 2021). However, when zooming out to a country range, like the study area of this work, not only areas in urban surroundings attract public attention and interaction. Areas such as Bavarian Alps and Saxony Switzerland National Park are also hot spots for tourists' attention. Hence, in the context of this research, when designing maps or map applications for purposes such as supporting travel planning, areas of interest (AOIs) can be defined as areas that expose to the public, have a relatively high frequency of visitations, and are noteworthy for a wide range of tourism-related applications. According to this definition, points of interest are specific locations that involve public interests.

Nonetheless, unlike federal states or cities that have clear administrative boundaries, an urban area of interest is a vague areal object, of which the borders are difficult to delineate. Liu, Yuan et al. (2010) proposed a point-set-based region (PSBR) model to approximate such kind of vague areal objects, which makes it possible to generate AOI boundaries out of a series of points of interest within the corresponding areas. However, AOIs generated by the PSBR model would have uncertainty, which is associated with the point pattern of POIs, more specifically, the density of POIs. Since the POI density varies dramatically among different regions, a relatively uniform accuracy of the generated

AOIs cannot be guaranteed. Besides, based on the mobility of tourists, especially during the initial travel planning stage, the precise boundaries of AOIs are either not necessary, since tourists can move conveniently within an area by vehicles and only an approximate range is needed, or should be combined with existed, known administrative boundaries to give the audience a definite perception of the geographic locations.

In this study, place of interest (POI) will be used as a generic term for area of interest and point of interest. While when discussing different methods used to summarise AOIs and extract points of interest, these two terms will be referred to explicitly.

2.2 Data Sources

2.2.1 Volunteered Geographic Information

Goodchild (2007) defines Volunteered Geographic Information (VGI) as “the tools to create, assemble, and disseminate geographic data provided voluntarily by individuals,” while Sui (2008) describes it as “the emergence of a new geography without geographers,” which reveals some characteristics of VGI: they are online, digital spatial data, usually created or produced by non-individuals and non-professionals, and the production of VGI is not charged.

Platforms, such as Wikimapia and OpenStreetMap, provide base maps, allow users to create, edit or update features by marking locations on base maps, assigning feature types, and complementing other information, use web browsers to visualize the spatial data, and let the public utilize the generated contents for further development or researches. Additional web applications that have embedded location-based services like Yelp where users can publish restaurant reviews and Dianping in China, which offers users chances to search and comment on local businesses, also enrich VGI content. The user-friendly interfaces, easy accessibilities, and open resources of these applications have been attracting more and more users from diverse groups to share, produce and contribute to VGI. The volume of existing, digital geographical data on a broader range of subjects has constantly been increasing using VGI tools (Elwood, 2008).

VGI has been influencing an expanding number of disciplines related to the domain of geographic information science. For example, by carrying a web-based survey among 202 high school and university students in Germany, Bartoschek and Keßler (2013) discovered that OpenStreetMap, the most frequently utilized VGI application in education, is being used in projects, regular courses, thesis work, and other relevant contexts by the participants. Education is also motivating them to use and contribute to VGI (Bartoschek & Keßler, 2013). In the field of disaster management, after reviewing and classifying 426 papers that explicitly mention using VGI in disaster management, Granell and Ostermann (2016) summarize five categories concerning the application of VGI regarding nat-

ural disasters which include crisis detection and prediction, crisis monitoring, etc. This shows that VGI, as a tool, is influencing different phases of crisis management. Besides, VGI also allows citizens to have a chance to actively participate in other fields, such as mapping, urban planning, tourism management (Ricker et al., 2013; Sun et al., 2013), land administration (Morero et al., 2018), etc.

2.2.2 Location-based Social Media

Kaplan and Haenlein (2010) set a general definition for social media: “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content”. However, they also state that there can be many sub-categories of social media within this definition. After a careful and clear classification assessed by social presence/media richness and self-presentation/self-disclosure, they conclude that social media can be categorized into six groups, as demonstrated in table 1. Nowadays, platforms such as Flickr, Instagram, Facebook, and Twitter all belong to social networking sites/social networking applications, which transmit a comparably medium amount of information and achieve a medium degree of social contact but can relatively represent more of individuals. In this study, the data used in analysis and visualization are mainly from this sub-category (social networking sites).

Table 1 Six sub-categories of social media (Kaplan & Haenlein, 2010)

		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/Self-disclosure	High	Blogs (e.g., The Moz Blog)	Social networking sites (e.g., Twitter)	Virtual social worlds (e.g., Second Life)
	Low	Collaborative projects (e.g., Wikipedia)	Content communities (e.g., Youtube)	Virtual game worlds (e.g., League of Legends)

It is popular for users on social networking applications to geo-tag their generated content sharing the current statuses. While location-based social media (LBSM) or location-based social network (LBSN) is more than just attaching an instant location or location history to the shared information; it also comprises a new social structure including indi-

viduals linked by the exact physical locations or location histories and geo-tagged contents (Zheng, 2011). Four trendy social networking applications: Flickr, Instagram, Facebook, and Twitter all provide geo-tag functions and own millions of monthly active users (MAU). Table 2 describes the abovementioned four social networking applications and lists their monthly active users according to the latest official data.

Table 2 Comparison of four social networking applications

Platform	Description	MAU
Flickr	Flickr is the photography revolution for sharing, storing, and organizing your photos in one of the largest worldwide photo communities.	60 million
Instagram	Instagram is a fast, beautiful, and fun way to share your life with friends and family.	1 billion
Facebook	Facebook enables users to consume content via the News Feed, chat with friends, create personal (and business) profile pages as well as share photos and videos, and join various groups.	2.89 billion
Twitter	Twitter is for people to see and talk about what is happening.	1.19 billion

Although only a tiny percentage of posts are geo-tagged, for example, according to Huang and Carley (2019), only 2.31% out of more than 40 million tweets collected from sample users are geo-tagged, due to the huge number of social media posts created every month, the volume of geo-tagged posts remains significant. Furthermore, due to the enormous data volume, most studies have relied on offline and historical social media data to conduct analysis and trials to date (Granell & Ostermann, 2016). Similarly, this study also implements data analysis and visualization based on the historical LBSM data which are introduced amply in chapter 3.

As shown in Table 2, these four social networking applications have different focuses on their own product positioning and thus attract different user groups. For AOI extraction and visualization on small-scale maps based on a large amount of data, using a combination of social media sources not only helps to increase the data volume but also covers non-overlapping user groups from different platforms, making the extracted AOIs more diverse and representative, rather than targeting a single user group.

Back to LBSM data itself, it has many similarities with VGI data: large volume, publicly visible, contains geographic information, and is generated and uploaded by non-specialists. But there is indeed a minor difference between them that VGI data is voluntarily uploaded by users and shared with the public without restrictions on use, but

LBSM data is uploaded by users for sharing purposes without being fully informed of the possible uses. However, in this paper, since data for analysis mainly come from LBSM and data for base maps that help to locate AOIs and POIs are from OpenStreetMap, I continue to use the term VGI although in an inclusive manner that data from LBSM are not explicitly volunteered.

2.2.3 Privacy Issues

The amount of data uploaded to social media is snowballing, while there is also a growing awareness of the value, potential, and risk of the personal data that we upload (Smith et al., 2012). Different social networking applications have adopted various measures and policies to address the increasing privacy concerns. For example, face recognition is used by Facebook to provide friend tagging suggestions based on friends who have already been tagged (Smith et al., 2012). Flickr only imports and displays GPS coordinates in a photo's EXIF header with the author's permission regarding location information protection. On Flickr, users can also create geofences and define the protected radius like in figure 2 for the exceptional locations that they think require particular or stricter privacy settings (Flickr, 2020).

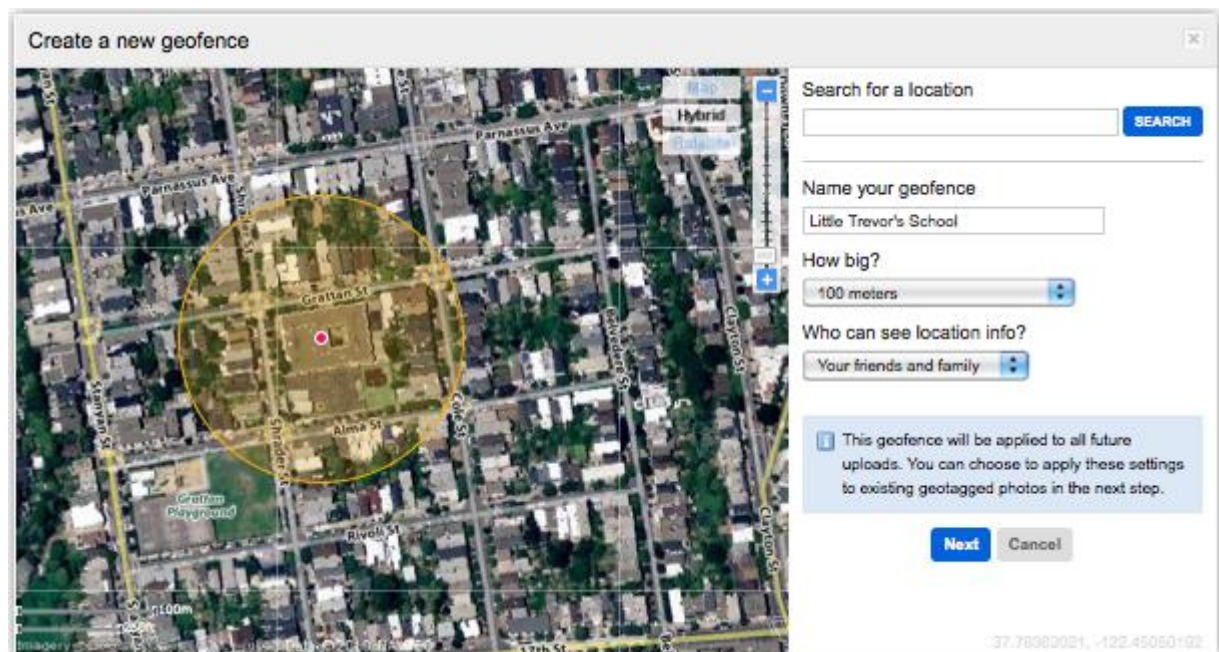


Figure 2 Create a new geofence on Flickr (Flickr blog, 2011)

For those posts whose privacy setting is “public,” since the post data is available on the Internet, it is likely that this data will be collected and used for other purposes, including, but not limited to, data analysis and visualization. Even if these users make their posts publicly visible at the point of upload, these posts are uploaded to social media for communication and sharing purposes without the users being fully aware of the poten-

tial for other uses. Therefore, it is inappropriate and against the wishes of the authors to not obfuscate, expose personal information about the author, or focus too much on the individual rather than on big data trends in any research that uses data from these public posts, especially for commercial purposes or third-party applications.

2.3 Data Processing and Visualization

2.3.1 HyperLogLog

Cardinality is the number of elements in a set, while HyperLogLog (HLL) is “a near-optimal probabilistic algorithm dedicated to estimating the cardinality of multisets” (Flajolet et al., 2007). HLL has been proved its advantages in user privacy protection, performance improvements, and a reduced storage need when dealing with LBSN data (Dunkel et al., 2020). While exploring and visualizing the potential phenomena and trends in a large volume of LBSN data, HLL makes it possible to isolate the data analysts and the original data ever since the data collection process. For example, once the topics that the data analysts are interested in are settled, the original data can be compressed into multiple parallel shards containing only the quantity information of the posts or users, which can be further aggregated.

The cardinality of a multiset can be precisely computed using raw data, which occupies massive storage space, especially when dealing with big data like VGI data, but the HyperLogLog algorithm relaxes this constraint by estimating the cardinality with a typical accuracy of 2%. Dunkel et al. (2020) used YFCC100m, which is “public-available and contains 100 million photos and videos from Flickr shared by 581,099 users under a Creative Commons License” as a use case for HLL. They compared raw data and HLL data performance from the input and output data size to processing time when the data is aggregated on 100 km grids, as in table 3. In this table, compared with raw data processing, HLL data processing has significant improvements in saving storage space and processing speed.

Table 3 Performance comparison for raw and HLL data processing (Dunkel et al., 2020)

Context	Raw Data	HLL Data
Input data size of comma-separated values (CSV)	2.5 GB	Explicit: 281 MB Sparse: 134 MB Full: 3.3 GB
Output data size, 100 km grid (CSV)	182.46 MB	19.80 MB
Processing time (Worldmap)	Post count: 7 min 13 s User count: 8 min 55 s	54.1 s (Post count, user count, user days)

	User days: 12 min 8 s	
Memory peak (Worldmap)	Post count: 15.4 GB User count: 15.5 GB User days: 19.3 GB	1.4 GB (Post count, user count, user days)
Benchmark data size (CSV)	/	10.61 MB (bins with user count ≥ 100)

HLL stores a structure of hashes instead of raw data. Figure 3 is an example of this structure, which is called a shard. This example is the user_hll of one grid containing three users, which is demonstrated in detail in subsection 3.1.2.

`\x128b7f9498ad68e108edc53cf199ed3f87f1125ed98858a38358c7`

Figure 3 An example of HLL shard structure

Another characteristic of HLL algorithm is that HLL allows lossless union operation on multiple sets, which makes calculating the number of distinct users (User Count) within each area after aggregation possible (see sections 3.3.1 and 3.3.2). Firstly, assume there are two regions that have been visited by X and Y tourists separately. The X tourists compose an HLL set A , while Y tourists compose an HLL set B , then the union of these two sets $A \cup B$ is comprised of all the tourists that have been to either one of the two regions. And the cardinality of the union set $|A \cup B|$ is the total number of the distinct tourists.

2.3.2 Density-based Clustering

The density-based clustering algorithm is an unsupervised clustering method designed to discover clusters with irregular shapes. The method considers the data set as a collection of several high-density clusters separated by low-density regions, groups the dense points that satisfy the conditions, and divides the combined high-density regions into clusters with connected densities and the largest set of points.

DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a pioneer density-based clustering algorithm that also works with databases containing even noise and outliers (Khan et al., 2014). DBSCAN has been independently implemented multiple times and is included in clustering toolkits (Schubert et al., 2017) such as scikit-learn (Scikit-learn, n.d.), R (R, n.d.), PostGIS (PostGIS, n.d.), etc. . It does not need a previous definition of the expected number of clusters in the data, but a set of parameters (ϵ , minPts) is required to characterize the sample (Crockett et al., 2017). The algorithm is implemented by drawing a circle called ϵ -neighborhood with ϵ (eps) as the radius for each data point as the center of the circle and then counting the points inside the circle, which is the density

value of that point. Then a density threshold MinPts is chosen, and the points in the circle that are greater than or equal to MinPts are high-density points called core points, and all the points within a ϵ radius of a core point are in one cluster and are direct density reachable to the core point. If any of these neighbor points becomes a core point again, their neighborhoods are included transitively called density reachable (Schubert et al., 2017). In contrast, the points in the circle that are smaller than MinPts and located in the circle of core point are low-density points called border points. Besides, the points that are neither core points nor border points are called noise points (Ester et al., 1996). Figure 3 depicts the cluster model of DBSCAN. The radius of the circles indicates the parameter ϵ and minPts is 4. In this illustration, all the red points, including A are core points, N is a noise point, and B and C are border points, while the arrows imply density reachabilities (Schubert et al., 2017).

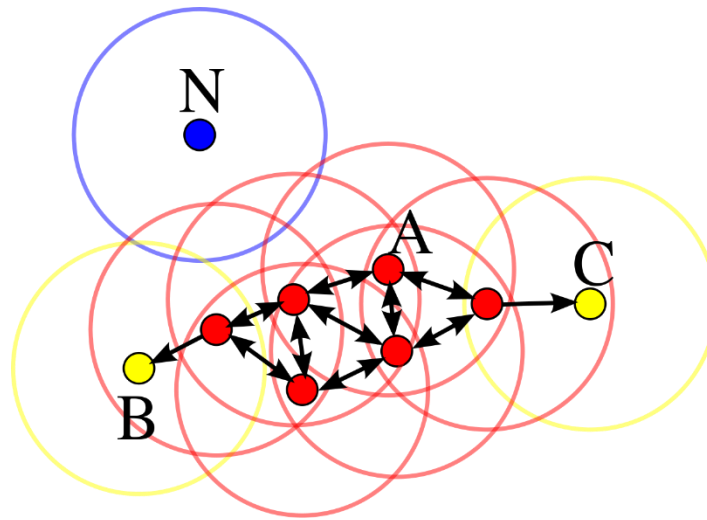


Figure 4 Illustration of the DBSCAN cluster model (Schubert et al., 2017)

Many researchers attempted to enhance the basic DBSCAN algorithm and proposed variations such as VDBSCAN (Varied DBSCAN), IDBSCAN (Integrated DBSCAN), KNNDBSCAN (K-nearest neighbors DBSCAN), etc. These algorithms improve DBSCAN by speeding up the clustering process, automating the computation of density threshold, and/or saving the main memory (Khan et al., 2014). However, Schubert et al. conclude that “the original DBSCAN algorithm, with effective indexes and reasonably chosen parameter values, performs competitively” by revisiting the statements in the SIGMOD 2015 article (Gan & Tao, 2015), conducting new runtime experiments, and doing new experimental evaluations. Furthermore, based on a good understanding of the given dataset for this work (see Flickr CCBY post dataset in subsection 3.1.3), two parameters (ϵ , minPts) required by DBSCAN can be adequately selected. Additionally, being available in various toolkits also increase ease of use which makes DBSCAN more practical to be applied in this study.

2.3.3 Interactive Web Map Application

Interactive web maps serve not only as navigation tools but also as visual analysis and visual communication tools. To build an interactive web map application, two perspectives should be considered: information and interaction. Web maps offer the possibility to include more information in a more concise and diverse way than paper maps. GIS organizes a variety of data that describe the natural world from different aspects into different layers and combines them on one map (Lobo et al., 2015). For example, the base map usually contains all the elements that exist objectively in reality, such as mountains, rivers, administrative boundaries, roads, etc. Therefore, the base map can also be considered as a context layer. Other layers that contain different types of data with geographic information either complement the base map or focus on a specific topic. For example, the layers that comprise subjective information of POIs belong to this, and they can be called focus layers. Geovisualization tools are used by professional cartographers as well as non-specialists in a variety of fields. By visualizing multiple sources of data with geo-locations in an exploratory manner and unifying them in a single representation, they can link these data and allow users to gain insight into potential patterns or anomalies (Lobo et al., 2015).

While cartographic interaction, which can be defined as “the dialogue between a human and a map mediated through a computing device” (Roth, 2013), offers users more possibilities in exploration and can help them gain geographic insight and support them in decision-making. When designing the cartographic interactions for a web map application, the input capabilities should be considered and fully utilized to achieve more flexible interactions. Classic keying devices such as keyboards and pointing devices such as mice and touchpads allow users to zoom in and out, hover, click on the map as well as type in input boxes. For the users of this tourist-targeted web map application that this work aims at, these possible interactions allow them to select and zoom in to the areas that they are interested in, avoid overwhelming them with information that is unrelated to their targets, and enable the audience to explore the map progressively.

3 Application

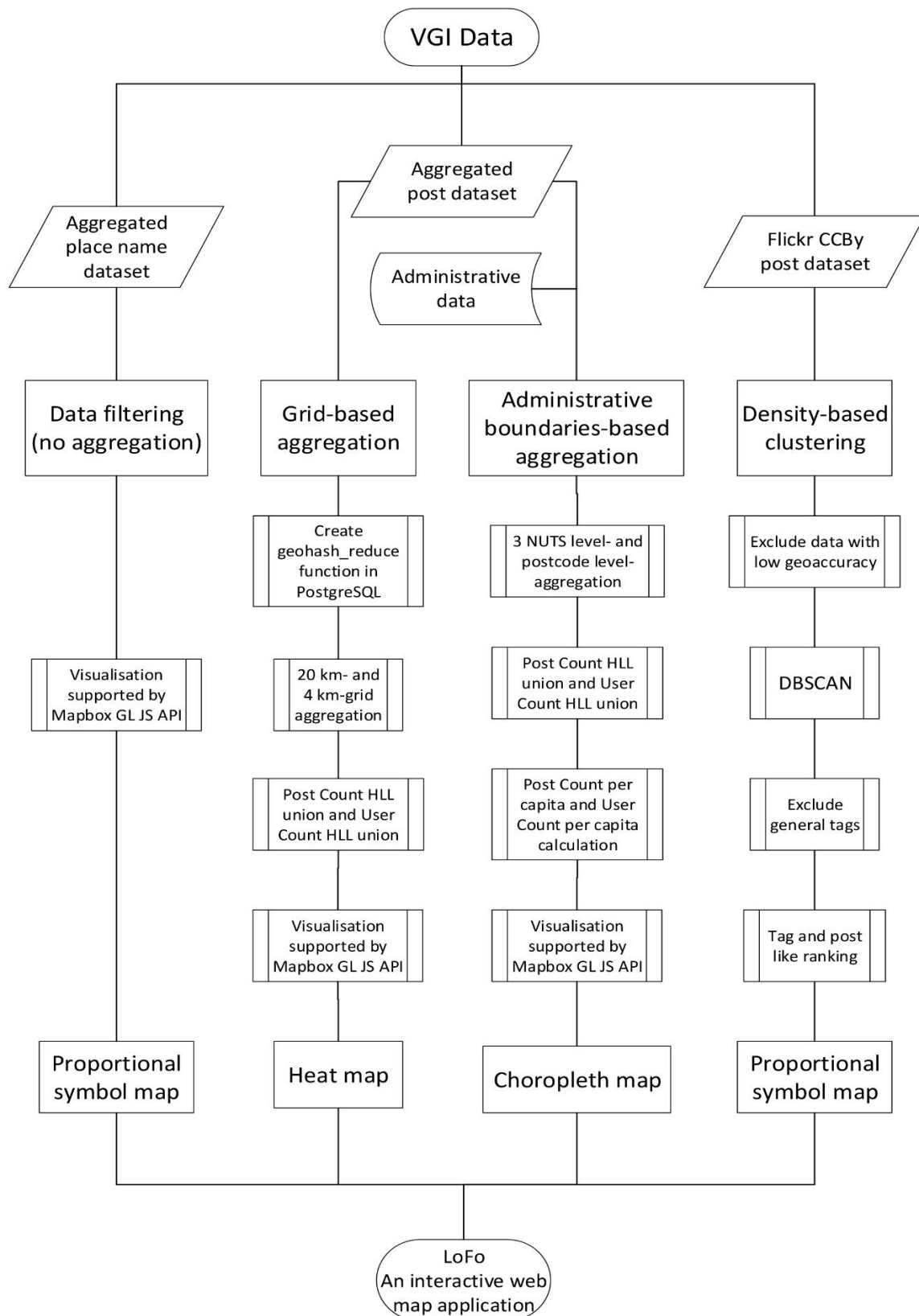


Figure 5 Workflow Diagram

Figure 5 illustrates the workflow of data processing and visualization from the original VGI data to the outputs, which is an interactive web map application combined with four maps. The following chapters 3 and 4 explain this workflow in detail and demonstrate the results.

3.1 LBSN Data

In this study, three VGI datasets in CSV (comma-separated values) format, one aggregated place name dataset, an aggregated post dataset, and one Flickr CCBY post dataset are used separately to explore different methods of extracting POIs. Table 4 describes these three datasets, which summarizes and compares their data sources, data volume, and applications in this study.

Table 4 Description of three VGI datasets

Dataset ID	Dataset	Source(s)	Data Volume	Application
1	Aggregated place name dataset	Instagram, Twitter, and Facebook	963012 (places)	Visualizing POIs on a large scale map with place names generated by users or social media applications
2	Aggregated post dataset	Flickr, Instagram, Facebook, and Twitter	40311403 (posts)	Summarizing and visualizing AOIs on small scale maps
3	Flickr CCBY post dataset	Flickr (2007-2021)	2864315 (posts)	Extracting and visualizing POIs on large scale map; information supplement

3.1.1 Aggregated place name dataset

The data in aggregated place name dataset includes 963012 place names in Germany which come from three out of four location-based social networking (LBSN) applications that have been introduced before: Instagram, Twitter, and Facebook. Table 5 demonstrates the data fields, their descriptions, data range when it is available, and whether the data field contains objective or subjective information based on if it reflects physically quantifiable properties of the environment or hardly quantifiable qualities. Table 6 compares the data from these three data sources, listing the number of records, postcount range, and data coverage.

Table 5 Data fields in aggregated place name dataset

Data field	Description	Range	Objective/Subjective information
origin_id	A unique origin_id helps to identify the data sources	1 (Instagram), 3 (Twitter), and 4 (Facebook)	(only about the data source, not about the environment)
place_guid	A globally unique identifier generated for a place	Globally unique	Subjective information
name	The name of a place that can be edited by users and automatically generated by LBSN	Within Germany	Subjective information
post_count	Number of posts that users have created and tagged at this place	[null] and 0 - 8122644	Subjective information
Location (Latitude and Longitude coordinates)	The geo-location of a place	Within Germany	Objective information

Table 6 Comparison of data from three social networking applications

	Instagram	Twitter	Facebook
Number of records	855747	69644	37621
Postcount range	0 - 8122644	[null]	[null]
Data Coverage	relatively complete	relatively outdated (~2016/17)	incomplete

The latitude and longitude coordinates have been processed with a Geohash length of 8 which means the geo-accuracy is approximately 19 meters. The total postcounts given are from the public APIs provided by Instagram. The postcounts of places retrieved from Facebook and Twitter are unknown in this dataset, making the data from these two LBSN applications not adaptable for extracting POIs since no standard can be used to judge

the popularity of each place. Additionally, too many places visualized on the same map layer can also be overwhelming for the audience. Thus, before visualizing this dataset, the data from Twitter and Facebook are excluded for achieving a better visualization result. Besides, to reduce the amount of data and speed up the loading of the web map, Berlin city has been selected as a study case to exhibit the visualization of local POIs based on this dataset, of which the result is illustrated in chapter 4.3.2.

3.1.2 Aggregated post dataset

The data in aggregated post dataset come from a combination of four LBSN applications aforementioned and have been preprocessed with aggregation on approximately 500-meter grids. Because of the pre-aggregation, this dataset containing more than 40 million posts but with only 317806 geo-points helps to speed up spatial aggregation and is suitable for summarizing and visualizing AOIs on small scale maps. With further aggregation on larger grids, some hidden attributes in the data are magnified and revealed.

Metrics (Dunkel et al., 2019) which can be used for visual analytics of this dataset, are also included, such as Post Count (PC), User Count (UC), and User Days (PUD). Under this context, the postcount of each record means the number of posts that were created and geo-tagged with the locations that are within this certain grid and have been aggregated to this geo-location. In comparison, usercount represents the number of users that have created a post or posts within the aggregated area. Compared with postcount, usercount reflects the real number of active users in a specific region and within the covered survey period, since there are users who created multiple posts with the same geo-tagged locations due to their fixed residences or offices, or for promotional purposes. When based on a large volume dataset, using usercount for tourism-purpose visual analytics can help to exclude the abovementioned influencing factors to a certain degree. Userdays in this dataset, which is the cumulation number of distinct user count each day in this area (Wood et al., 2013), is not included in the further aggregation and is not used for visual analytics in this study.

Additionally, when doing further aggregation, even though postcount values can be directly added, usercount has to be counted distinctly since there can be the same users counted in adjacent grids that are aggregated into one grid. Thus, HyperLogLog is applied to these three metrics to solve this problem. Table 7 lists the data fields, their descriptions, data range when it is available, and whether the data field contains objective or subjective information.

Table 7 Description of data fields in aggregated post dataset

Data Field	Description	Range	Objective/Subjective information
Location (Latitude and Longitude coordinates)	The geo-location of a point that represents the corresponding grid	Germany	Objective information
post_hll	HyperLogLog format of posts in this grid	None	Subjective information
date_hll	HyperLogLog format of userdays in this grid	None	Subjective information
user_hll	HyperLogLog format of users in this grid	None	Subjective information
postcount	Cardinality of post_hll	1-484952	Subjective information
userdays	Cardinality of data_hll	0-280001	Subjective information
usercount	Cardinality of user_hll	1-197659	Subjective information

3.1.3 Flickr CCBY post dataset

Flickr CCBY post dataset contains Flickr Creative Commons images between 2007 and June 2021 in Germany, which are 2864315 posts after preprocessing. When users post photos on Flickr, they can choose the photo licenses out of 11 available categories for the photos. Flickr services assign license ids from 0 to 10 to these 11 types of licenses (Flickr Services, n.d.). During the preprocessing of this dataset, the images that are all rights reserved and belong to United States government work are excluded by filtering by their license id, post data used are all allowed under this study context.

Compared with other social networking applications, users on Flickr appear to pay more attention to the contents and quality of photos. In this dataset, Tags, emojis, post titles, post bodies, and images help present multi-dimensional information about the geo-tagged locations. Therefore, data from Flickr are more suitable for extracting local POIs and implementing and enhancing POI information. In this study, the thumbnails of photos and tags are selected for POI information enhancement.

In total, there are 20 data fields in this dataset. Post_guid and user_guid are unique identifiers generated for referencing the posts and users, and they are not utilized in this study to better protect users' privacy. Table 8 omits some data fields that are not used in data aggregation and data visualization, such as post_filter, emoji, post_title, post_body, etc., and lists the data fields, their descriptions, data range when it is available, and whether the data field contains objective or subjective information.

Post_content_license is only an attribute of posts that were set by the authors when they were created, but it still can reflect the willingness degree of how the authors share the photos about the places to the public. Therefore, it can be categorized as subjective information. Additionally, the coordinates of each post is generated by authors marking on the map. Therefore, geographic coordinates are actually subjective to the users, but longitude and latitude are still objective descriptions of the marked locations and can be classified as objective. In contrast, for each post, post_url is the objective address of the post on Internet. But a post_url actually infers to the post and contains subjective information.

Table 8 Description of data fields in Flickr CCBY post dataset

Data Field	Description	Range	Objective /Subjective information
origin_id	A unique origin_id helps to identify the data sources	2 (Flickr)	(only about data source)
Location (Latitude and Longitude coordinates)	The geo-location of a post	Germany	Objective information
post_thumbnail_url	Url to the public thumbnail of this post (usually this will only be available for posts of type IMAGE)	[null]	Subjective information
post_views_count	Number of times this post has been viewed by other users	0-312269	Subjective information
post_like_count	Number of times this post has been liked by other users	0-484952	Subjective information
post_url	Url to the original post	[null]	Subjective information
tags	List of tags assigned to the post	[null]	Subjective information

post_geoaccuracy	The highest location accuracy available for this post	['latlng', 'place', 'city', 'country'.]	Objective information
post_content_license	An integer for specifying the license of the post	1,2,3,5,6,7,9,10	Subjective information

3.2 Administrative data

Combining vague areas like AOIs with existing, known administrative boundaries can help users associate the AOIs to different levels of administrative regions that have defined geographic locations and scopes. Since this work aims to visualize data on maps of multiple scales and when the display area size stays the same, the area range that the viewers can see varies along with map scales, a hierarchical system of administrative regions with graded changes in the area is needed for aggregation. Nomenclature of Territorial Units for Statistics (NUTS) is such a hierarchical system that establishes a hierarchy of three NUTS levels in each EU member country and the UK for statistical, social, economic, and political purposes (Eurostat, n.d.).

In Germany, the three NUTS levels are states (NUTS1, German: Bundesland), government regions (NUTS2, German: Regierungsbereich, or Direktionsbezirke), and Districts (NUTS3, German: Kreis), with generally 3 to 7 million residents, 800 000 and 3 million residents, and 150 000 to 800 000 residents separately (Destatis, n.d.). A code is assigned to each EU region for unambiguous identification, and the length of the code can be 3, 4, or 5 digits depending on the hierarchical level. For example, on the NUTS1 level, Bavaria has code DE2, while Oberbayern on NUTS2 level has code DE21 and Munich on NUTS3 level has a code of DE212.

The shapefile format dataset of German NUTS boundaries used in this study come from the Federal Agency for Cartography and Geodesy (German: Bundesamt für Kartographie und Geodäsie) and consists of the geometry (.shp file), geometry index (.shx file), projection (.prj file), attributes (.dbf file), and character set (.cpg file) for each NUTS level in UTF-8 (Unicode) character encoding. For each NUTS level, the location, shape, and attributes, including NUTS code, NUTS name, and population, are necessary for further aggregation. The NUTS level used for aggregation increases along with the map scale.

Within NUTS3 districts, there are also local administrative units, while for Germany, they are called municipalities (German: Gemeinden). However, in real-life scenarios, municipalities are more applicable for political purposes, such as setting a citizen's office (German: Bürgerbüro) in each municipality. When the visual analytics is targeted to tourists, the postcode is more practical since tourists tend to know the postcode of the place where they stay overnight, and the postcode can conveniently help them to locate the

area within walking distance of their accommodation. Therefore, areas divided according to postcode are used to aggregate data on a larger scale map, allowing tourists to have a more detailed view of AOIs.

The shapefile format dataset containing the shape, location, and population of post-code areas within Germany comes from SUCHE-POSTLEITZAHL.ORG (SUCHE-POSTLEITZAHL.ORG, 2020), of which the row data source is OpenStreetMap and the population data source is Federal state statistical offices (German: Statistische Ämter des Bundes und der Länder).

3.3 Data Aggregation

3.3.1 Grid-based Aggregation

Aggregated post dataset is aggregated based on grids in this subsection. First, this dataset in CSV format needs to be imported into PostgreSQL (postgresql.org) so that further aggregation and processing can proceed with the help of PostGIS that allows running spatial queries in SQL. In order to process the HyperLogLog structure data, a Postgres module (Citusdata/postgresql-hll, n.d.) that introduces the new data type “hll” into PostgreSQL needs to be run by connecting to a PostgreSQL Docker Container “hlldb” (docker.com), which has this postgresql-hll extension installed. A table named `data_original` specifying the data field names and data types is created in the database system to store the aggregated post dataset. Besides latitude and longitude that are assigned as double-precision, the `data_original` table also contains another column `the_geom`, which is assigned with geometry data type representing the points in planar coordinate systems.

A spatial reference identifier (SRID) is “a unique identifier associated with a specific coordinate system, tolerance, and resolution” (ArcGIS Desktop, n.d.). In this study, the geometry column of all the spatial data that are processed in PostgreSQL is set as SRID equal to 4326, meaning that the spatial data use latitude and longitude coordinates that are defined in the WGS84 standard. Unifying the spatial reference system makes it possible to run spatial queries on geometry columns from different spatial datasets.

CSV format dataset is imported via `psql` that is a command-line interface for interacting with PostgreSQL. And code 1 shows how the dataset is imported using `/copy` command. In this command, “`csv`” tells that the file being copied is in CSV format, while “`header`” tells that the headers at the top of the file should be included.

```
hlldb=# \copy data_original(latitude, longitude, posthll, datehll, userhll, postcount, userdays, usercount) FROM 'THE CSV FILE FULL PATH' DELIMITERS ',' CSV HEADER;
```

Code 1 Command that imports CSV format dataset

After this process is finished, the interface prints “COPY 317806” meaning 317,806 records have been successfully imported in table data_original. Figure 6 shows the data in Berlin viewed in PostGIS geometry viewer.

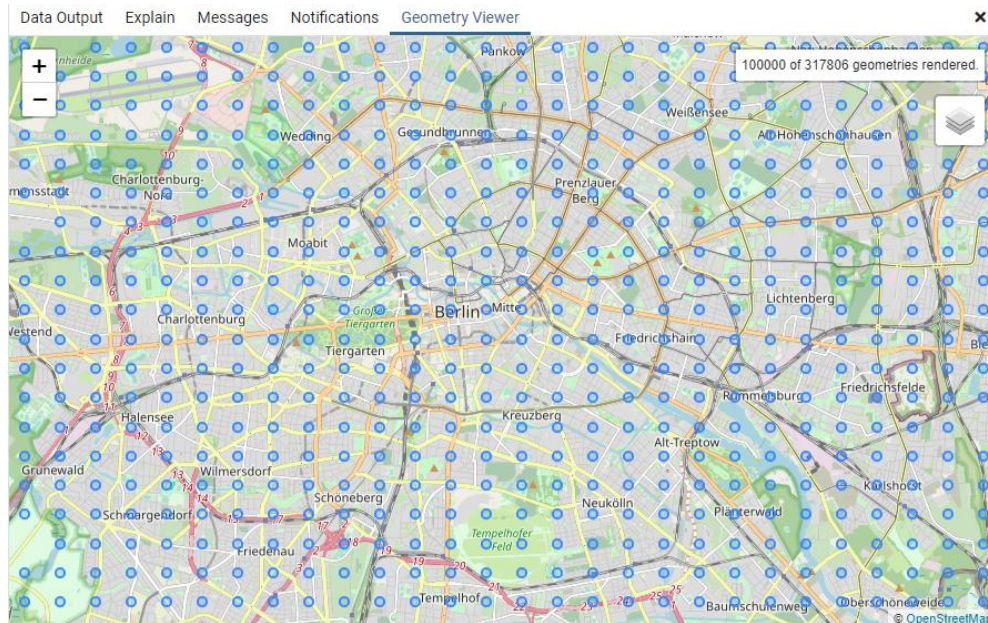


Figure 6 Original aggregated post data within Berlin

To further aggregate the data so that each point can represent the data of a larger grid, a geohash_reduce function is applied to aggregate latitude and longitude coordinates. AS a geocode system that encodes a two-dimensional latitude and longitude coordinate into a one-dimensional string composed of letters and digits, Geohash divides the earth's surface into buckets of grid shape. Table 9 lists the Geohash length in digits and the corresponding grid sizes.

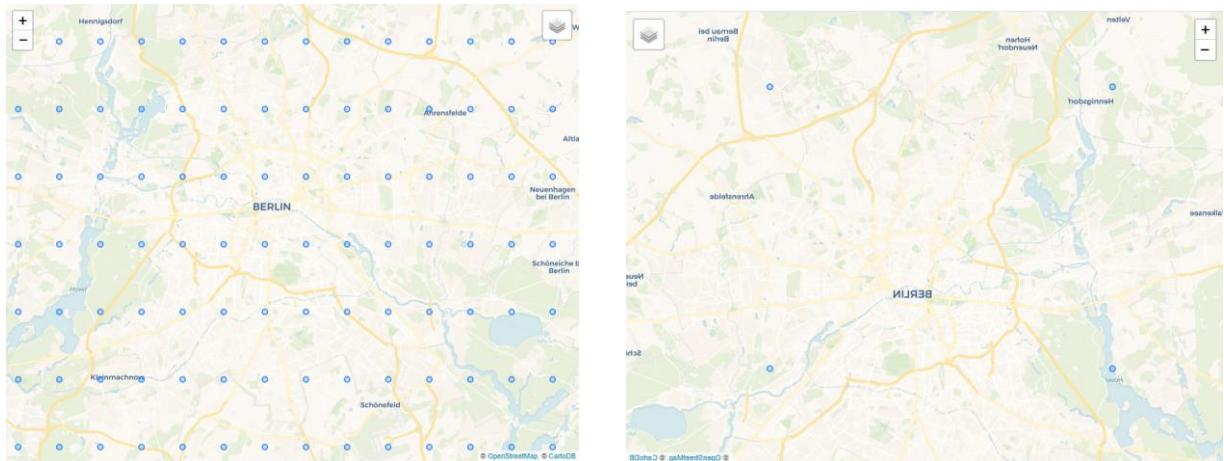
Table 9 Geohash length and distance of adjacent cell in meters (VGIscience, n.d.)

Geohash length (number of digits)	Distance of adjacent cell (m)
1	5003530
2	625441
3	123264
4	19545
5	3803
6	610
7	118
8	19
9	3.71
10	0.6

The `geohash_reduce` function (see Appendix A) asks the point geometry columns and the length of Geohash as input and returns the point geometries. The input point geometry is first converted to a Geohash of a specified number of digits using `ST_GeoHash` function provided by PostGIS, and then it is converted back to geometry data type by function `ST_PointFromGeoHash`. Four adjacent points with 7-digit Geohash such as `gbsuv6e`, `gbsuv6k`, `gbsuv6s`, `gbsuv67` that have the same first five digits can be aggregated to the same point of which the Geohash value is “`gbsuv`” using the `geohash_reduce` function.

Since this study aims to produce multi-scale maps on which the users can be informed with different degrees of details about the AOIs, the data are aggregated into two different sizes so that the results can be visualized on two map scales. Thus, the Geohash of point coordinates is reduced to 4 or 5 digits so that the points can be aggregated on approximately 20 km grids or 4 km grids (code see Appendix A). Two tables, “`point_20km`” and “`point_4km`,” are created to store the data after aggregation and comprise 884 and 25007 records separately.

Meanwhile, `user_hll` and `post_hll` as two hll type columns are also unioned together with the geometry column aggregation using “`hll_union_agg`” function. The distinct number of posts and users for the new grid can be derived by calculating the cardinalities of corresponding `user_hll` and `post_hll` (code see Appendix A). Figure 7 displays the data within Berlin after aggregation in geometry viewer. Since both tables contain multiple records, they are exported as FeatureCollection objects containing multiple points in .geojson files with the help of “`ST_ASGeoJSON`” function. Two geojson files can be used as data sources for interactive visualization in chapter 4.



(a) Data aggregated on 4 km-grids

(b) Data aggregated on 20 km-grids

Figure 7 Data within Berlin after aggregation

3.3.2 Administrative Boundaries-based Aggregation

Aggregated post dataset is aggregated based on NUTS boundaries as well. Shapefiles of the administrative data mentioned in subsection 3.2 are imported into the database via PostGIS Bundle 3 for PostgreSQL x64 13 Shapefile and DBF Loader Exporter. VGI data are further aggregated on five levels, including a country-level which provides a summary of the dataset, three NUTS levels, and postcode areas to visualize the results on various map scales. The. In PostgreSQL, five tables store the name, code, population, and geometry of each area (Germany, federal states, government regions, districts, and postcode areas) separately after the shapefile import.

The same table, "data_original," is directly employed for the aggregation. Spatial queries (code see Appendix B) are run on a data_original table and one administrative data table each time to achieve the aggregation for each level. For example, on NUTS3 level, during the aggregation, user_hll and post_hll of the points that are within one district are unioned. Therefore, the cardinalities of hll sets are filled in "postcount" and "usercount" columns.

Since the population and area of each federal state, government region, or district vary considerably, the population difference highly influences the total number of posts and users of each area. Post_per_capita and user_per_capita are implemented instead of postcount and usercount to reduce the impact of population on AOI popularity estimation, and they are calculated during the aggregation process.

Administrative boundaries-based aggregation produces five tables containing NUTS code or postcode, area name, postcount, usercount, post_per_capita, user_per_capita, and geometry of the areas on five separate levels. Figure 8 illustrates the data in the geometry viewer, which are the VGI data within Germany after aggregation on three NUTS levels. Each of these tables can be exported via PostGIS Bundle 3 for PostgreSQL x64 13 Shapefile and DBF Loader Exporter as shapefiles that can be used for interactive visualization.

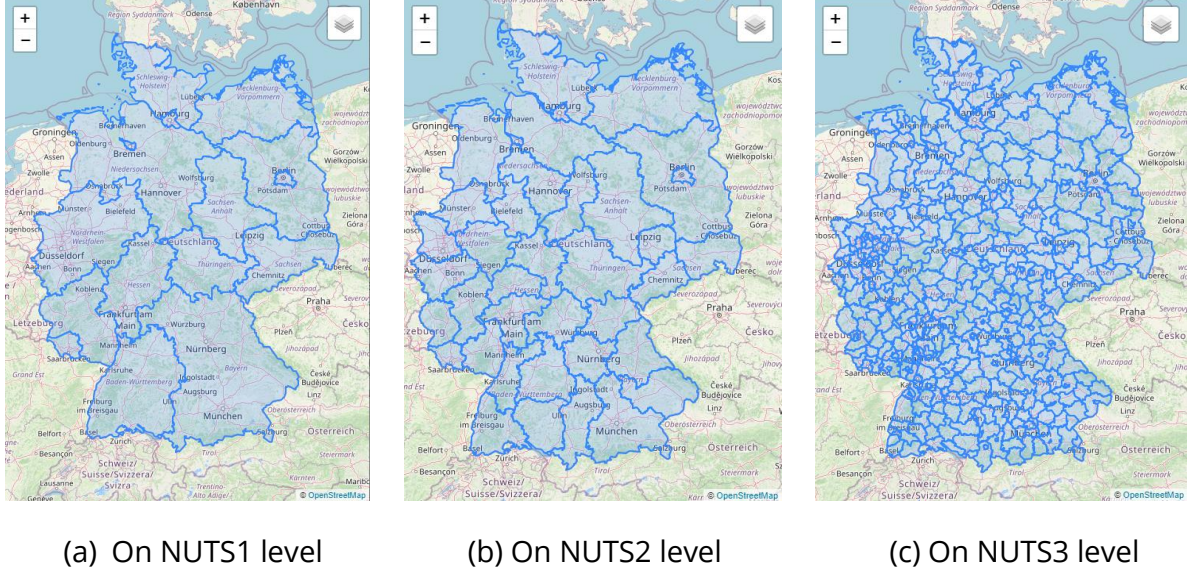


Figure 8 Data within Germany after NUTS boundaries-based aggregation

3.3.3 Density-based Data Clustering

For extracting POIs on a local level such as famous buildings, parks, or bridges, grid-based aggregation and administrative boundaries-based aggregation are not adequately detailed. With urban context, geo-accuracy higher than 30 meters is required for POI extraction, which makes DBSCAN (density-based spatial clustering of applications with noise) a better aggregation method because of its adjustable input parameters ϵ (eps) and minPts. In this work, Dresden city is selected as a study case for local level POI extraction.

Similar to subsection 3.3.1, a table named flickr_original is first created in the database containing all the data fields with appropriate data types assigned. Flickr CCBY post dataset is imported via the /copy command in psql so that queries can be run on this dataset in PostgreSQL. Only the data inside Dresden are kept, and they are further filtered and stored in table “flickr_dresden” due to data with low geo-accuracy, including city and country post_geoaccuracy. The data with a post_geoaccuracy as place and latlng are suitable for within-city level aggregation.

The extraction is done with the help of ST_ClusterDBSCAN provided by PostGIS. Code 2 does the data clustering when ϵ (eps) is 10 meters, and minPts is 30. A materialized view “cluster10_30” consists of cluster id, the number of points that belong to the cluster, and the centroid of the geometry collection of a cluster.

```

CREATE MATERIALIZED VIEW mviews.flickr10_30 AS
SELECT post_title,
       ST_ClusterDBSCAN(geom_m, eps := 10, minpoints := 30) OVER () AS cid,
       geom
FROM flickr_dresden;

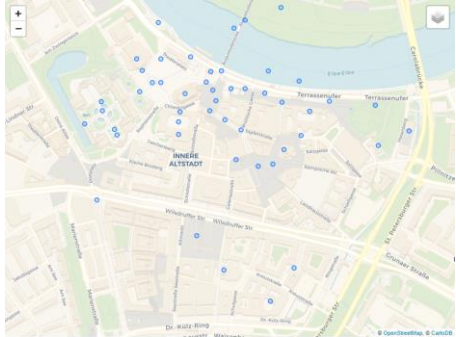
CREATE MATERIALIZED VIEW mviews.cluster10_30 AS
SELECT cid,
       COUNT(cid) AS num,
       ST_Centroid(ST_Collect(geom)) AS geom
FROM mviews.flickr10_30
GROUP BY cid

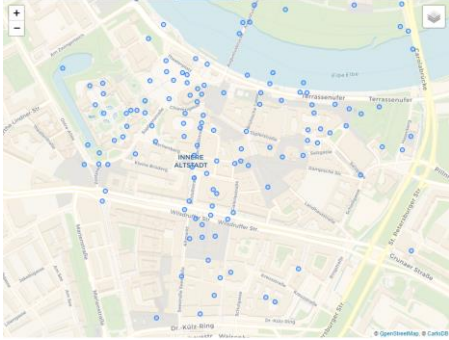
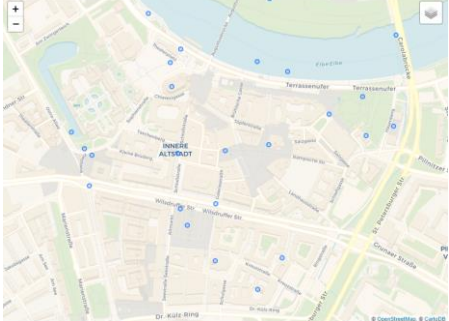
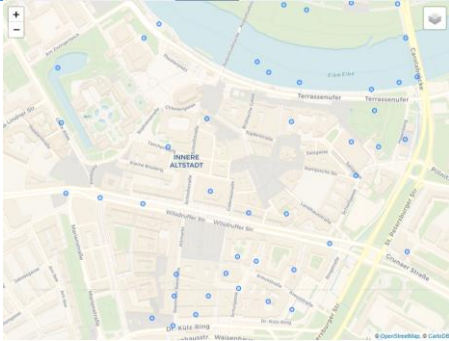

```

Code 2 Density-based data clustering in Dresden (eps=10, minPts=30)

Two required parameters by DBSCAN are selected with a good understanding of the data. As a local-level POI extraction is aimed to be done with DBSCAN, ϵ (eps) which is not larger than 50 meters appears to be appropriate so that specific popular photo-shooting spots, such as a market, and a museum, can be identified. At the same time, minPts highly varies along with the VGI data volume within particular cities. For this reason, several sets of parameters are chosen for the experiment and obtaining the comparably optimal parameters. The outcome in the area around Innere Altstadt in Dresden is selected as evaluation criteria. Some tourist attractions within this area, such as Frauenkirche, Altmarkt, and Zwinger, should be identified if the parameters are adequately picked. Table 10 illustrates the comparison.

Table 10 Comparison of results after density-based clustering

(eps in meters, minPts)	Number of clusters	Results in Innere Altstadt	Comments
(10, 30)	133		Frauenkirche, Altmarkt, and Zwinger are clearly identified; moderate amount of clusters.

(10, 10)	419		Frauenkirche, Altmarkt, and Zwinger can be identified, but the excessive amount of clusters is overwhelming.
(20, 30)	120		Zwinger is not marked with cluster centroid(s); moderate amount of clusters.
(20, 10)	327		Frauenkirche, Altmarkt, and Zwinger are not marked with cluster centroid(s).
(50, 30)	90		Frauenkirche, Altmarkt, and Zwinger are not marked with cluster centroid(s); too few marked places (clusters) within Innere Altstadt.

With parameters $\text{eps}=10$, $\text{minPts}=30$, DBSCAN produces better results than using other sets of parameters in table 10, and they are chosen as the data source for subjective information processing and further interactive visualization

3.3.4 Subjective Information Processing

Subjective information such as postcount and usercount can be calculated during the process of aggregation, which helps to weigh the popularity of each AOI or POI. However,

besides the popularity information, tourists are also interested in the “content” of a POI. In other words, during the trip-planning or searching for photo-shooting spots stage, the tourist or photographers would like to know what they can see at specific places. Since the users tend to attach tags that are highly related to the topics of the photos that they post and they pay more attention to the photo contents and quality, tags and thumbnails of photos from Flickr CCBY post dataset are utilized and processed for POI information enhancement and supplement.

```
CREATE TABLE mdata.piclink AS
SELECT cid,
       post_thumbnail_url,
       post_url
FROM
  (SELECT cid,
         post_thumbnail_url,
         post_url,
         ROW_NUMBER() over(PARTITION BY cid
                           ORDER BY post_like_count DESC) AS ranks
   FROM mdata.flickr10_30) AS a
WHERE a.ranks=1
AND cid IS NOT NULL
```

Code 3 Obtaining the most-liked post within each cluster

After the density-based data clustering, points within Dresden are aggregated into 133 clusters. Due to the limitation of web map display space, not all the photo thumbnails can be included, and it is also necessary to do so since a combination of popular tags and the most liked photo should be enough to provide an overview of a place. Code 3 ranks all the posts within one cluster by the `post_like_count` values and gets the `post_thumbnail_url` and `post_url` of the most liked post of each cluster.

For the tags, only five tags that occur the most frequently in the posts within each cluster are selected to apply in the interactive web map. Picking relatively popular tags helps to filter the “personal” tags that may not sufficiently represent the places but are more individual-related.

Additionally, when dealing with the posts in Dresden, tags such as Dresden, Germany, Sachsen, Saxony, Deutschland, and empty tags are excluded using “WHERE tag NOT IN ('dresden', 'germany', 'sachsen', 'saxony', 'deutschland', '');” when ranking the tags. Posts on Flickr tend to carry multiple tags that can be slightly related or highly related to the photo-shooting spot. Including some general tags as abovementioned would sieve out those comparably more related tags that we need.

For instance, the top five tags of a cluster located in Altmarkt after the tag ranking and filtering (see Appendix C) are altmarkt, innenstadt, baustelle, weihnachtsmarkt, and panorama. “Altmarkt” and “innenstadt” are related to the place name and location, while “baustelle” and “weihnachtsmarkt” reveal that there has been construction and it is a place where Christmas market is held.

4 Interactive Visualisation

4.1 User Interface

A web application based on HTML5, CSS3, and JavaScript is developed to help visualize the VGI data that has been aggregated and achieve interaction with targeted users. The interaction possibilities will be elaborated in the following sub-sections. This application, which is named as LoFo, is shown in figure 9 and is mainly composed of two parts: navigation bar on the top and map display area only exhibiting the visualization within the study area of this work (Germany).

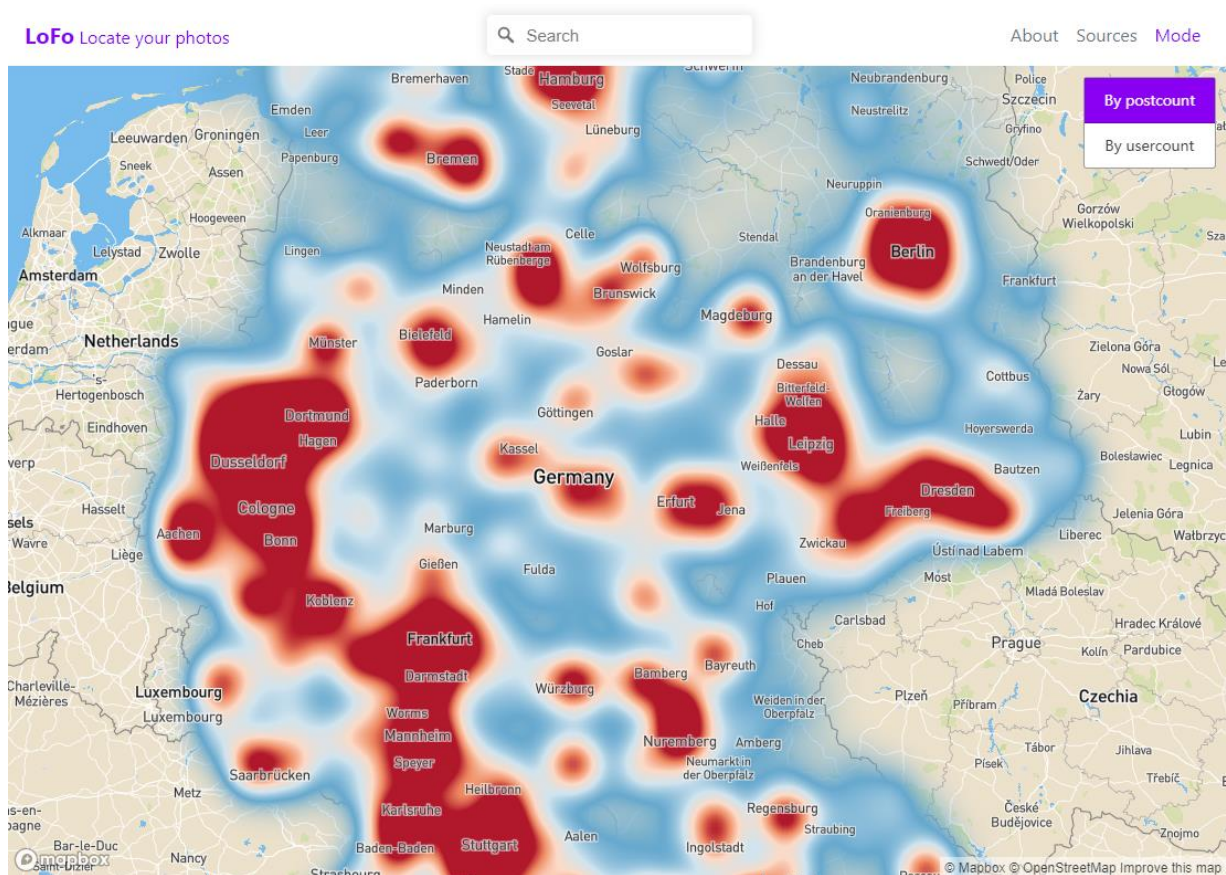


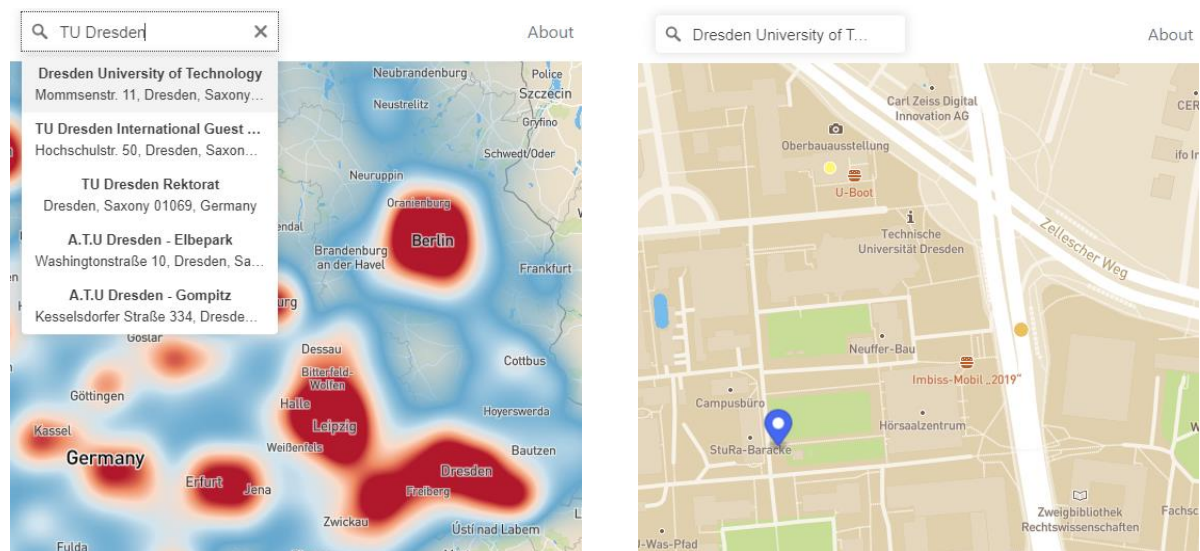
Figure 9 The user interface of LoFo including a navigation bar and a map display area

A purple theme color (color reference #8900F2) is applied throughout this interface. Choosing this theme color helps users to reinforce the impression of the brand or application and helps to achieve consistency and coherence across sections and pages of the application. Highlighting individual buttons with a theme color also draws the user's attention and guides the user's interactions. For example, in the navigation bar, which locates at the top of the whole interface and serves as a control row linking users to different sections of this web application, the theme color highlights the logo section in the top

left corner, which always navigates users to the main page via clicking and the “Mode” button in the button set which locates in the right top corner. In this context, it impresses users with a brief introduction of this web application and attracts users to click the “Mode” button that allows them to switch the mode and explore other types of visualization.

The other two buttons in the button set are “About” and “Sources.” The “About” button links users to an information page, which introduces the idea and functions of this web application. Developer contact information is also displayed on this page. The “Sources” button navigates users to a data source page which informs the audience about from which social media site data are collected and the privacy protection measures.

Another part besides the logo section and button set is a search box in the middle of the navigation bar, which is a geocoding control that works for all the web maps included in this web application. The search bar loads the mapbox-gl-geocoder plugin into LoFo and geocodes users’ input. As demonstrated in figure 10, it enables users to search the places that they have interests. For example, in figure 10(a), when users type in TU Dresden in the search box, related possible results will be listed below the search bar. Once one of the results is selected, it will move and zoom in the camera directly to the destinations, as shown in figure 10(b). This feature uses the Mapbox Geocoding API and is based on OpenStreetMap data and helps users locate their areas of interest or points of interest more quickly and precisely.



(a) Results returned by the search box

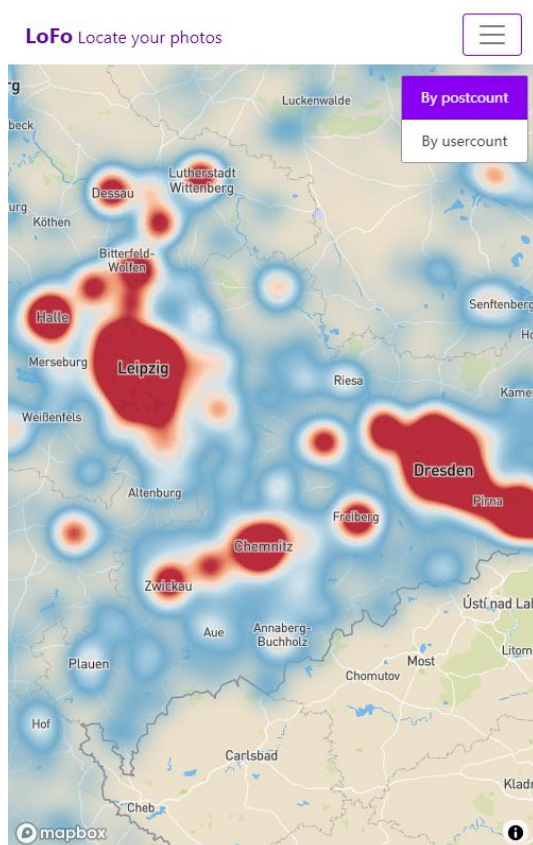
(b) The map zooming in to the result

Figure 10 Diagram of how the search box works

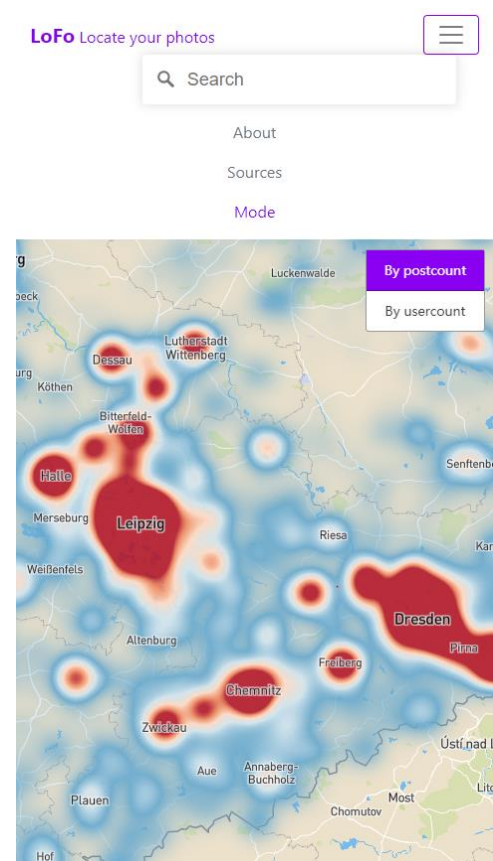
In the map display area below the navigation bar, data from after aggregation are visualized based on map styles and functions provided by Mapbox GL JS API and base map data from OpenStreetMap. They serve the audience as interactive web maps with control

buttons, an information window, and pop-up windows. The following sections in this chapter describe these four types of web maps in detail.

LoFo is also a mobile-compatible web application. When users open LoFo on extra small devices (e.g., smartphones) which are smaller than 768px wide, the search box and the button will be hidden and embedded together in a hamburger menu located in the right top corner, as displayed in figure 11(a). In figure 11(b), they appear below the navigation bar after clicking on the hamburger menu. Additionally, after testing in various browsers, LoFo is compatible with all the popular web browsers such as Google Chrome, Firefox, Safari, and Microsoft Edge. Being mobile-compatible and cross-browser compatible makes LoFo better serve the targeted users. Users like tourists, photographers, or people who are interested in exploring POIs can get an overview of the POIs all over Germany or check specific popular locations within cities. Users can conveniently use any devices that can connect to the Internet to check where other tourists/photographers took pictures and what spots attracted them most. This application will hopefully help with users' decision-making -- where to visit and where to take photos.



(a) The navigation bar on a mobile device



(b) The hamburger menu after clicking

Figure 11 The user interface of LoFo on a mobile device

4.2 Visualization for Overview

4.2.1 AOI Heatmap

As discussed in sub-chapter 2.2, AOIs are vague areal objects which usually do not have clear boundaries. AOIs are generated by a large number of points of interest that carry popularity information and are close to each other or overlapping. A heatmap uses a system of color-coding and the principle that grayscale can be superimposed to represent those quantified popularity values such as postcount and usercount. Therefore, a heatmap is a perfect tool for visualizing AOIs, which allows the viewers to zoom out from points of interest, gives them an overview of AOIs and their approximate locations, and lets them explore the potential information and pattern hidden with the point set.

While users can switch metrics between postcount and usercount, there are in total four heatmaps varying along with zoom levels in LoFo. Table 11 describes zoom levels in Mapbox based on the data from OpenStreetMap and their different degree of detail. As users zoom in, one vector tile covers a smaller range of areas but together with more information. Figure 9 shows the map using postcount as a metric with a zoom level at 4 in the initial state of the web page. AOIs such as the Ruhr area, Hamburg, and Berlin can be identified at this zoom level. From zoom level 0 to a maximum zoom level 7, data after aggregation on Geohash length of 4 is applied for the visualization.

Table 11 Zoom levels (Mapbox, 2021)

At zoom level	Number of tiles	You can see
0	1	The Earth
3	64	A continent
4	256	Large islands
6	4096	Large rivers
10	1048576	Large roads
15	1073741824	Buildings

From a minimum zoom level 7 to zoom level 9, data after aggregation on Geohash length of 5 is utilized. Gradient effects are used to bridge zoom levels more smoothly. Figure 12 is a schematic diagram displaying the result after zooming in to the area around Saxony. AOIs such as Dresden, Leipzig, and Chemnitz have relatively high postcounts with the whole federal state.

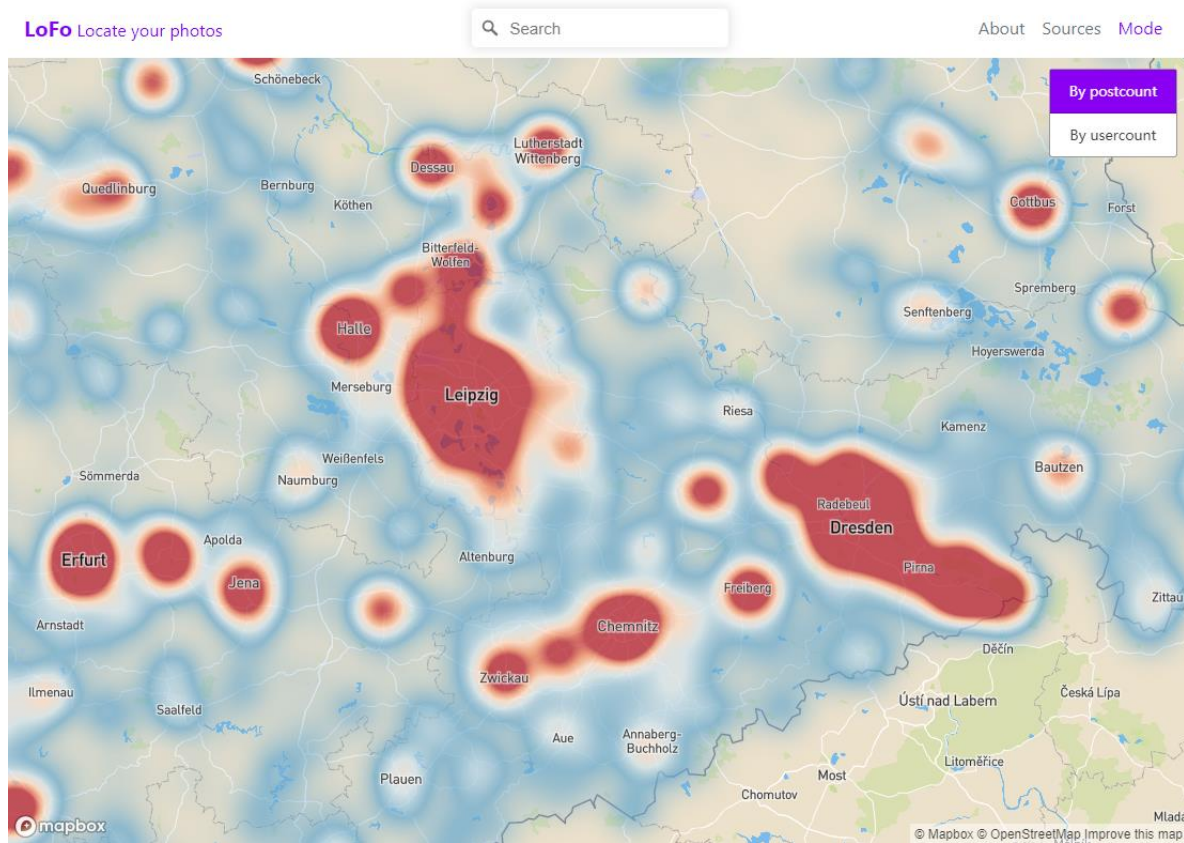


Figure 12 Heatmap displaying the region around Saxony (at zoom level 8)

As shown in figure 13 and figure 14, which zooming into southeast Germany, after selecting a metric by clicking the toggle buttons in the right top corner of the map display area, both maps with postcount or usercount as the metric exhibit similar results from zoom level 7 to zoom level 9. Cities such as Ulm, Augsburg, Rosenheim, and Munich have relatively higher postcount and usercount per Geohash grid and are illustrated as hotspots with dark red color on the heatmap. Besides, the two maps both allow users to identify the AOIs around Füssen where Schloß Neuschwanstein is located and Garmisch-Partenkirchen, with a large number of posts and active users. This also echos to the introduction in the study area section. Another AOI on the right side of Rosenheim, which can be found on both maps, is the area around Lake Chiemsee, where many tourists take trips or enjoy vacations all year long.

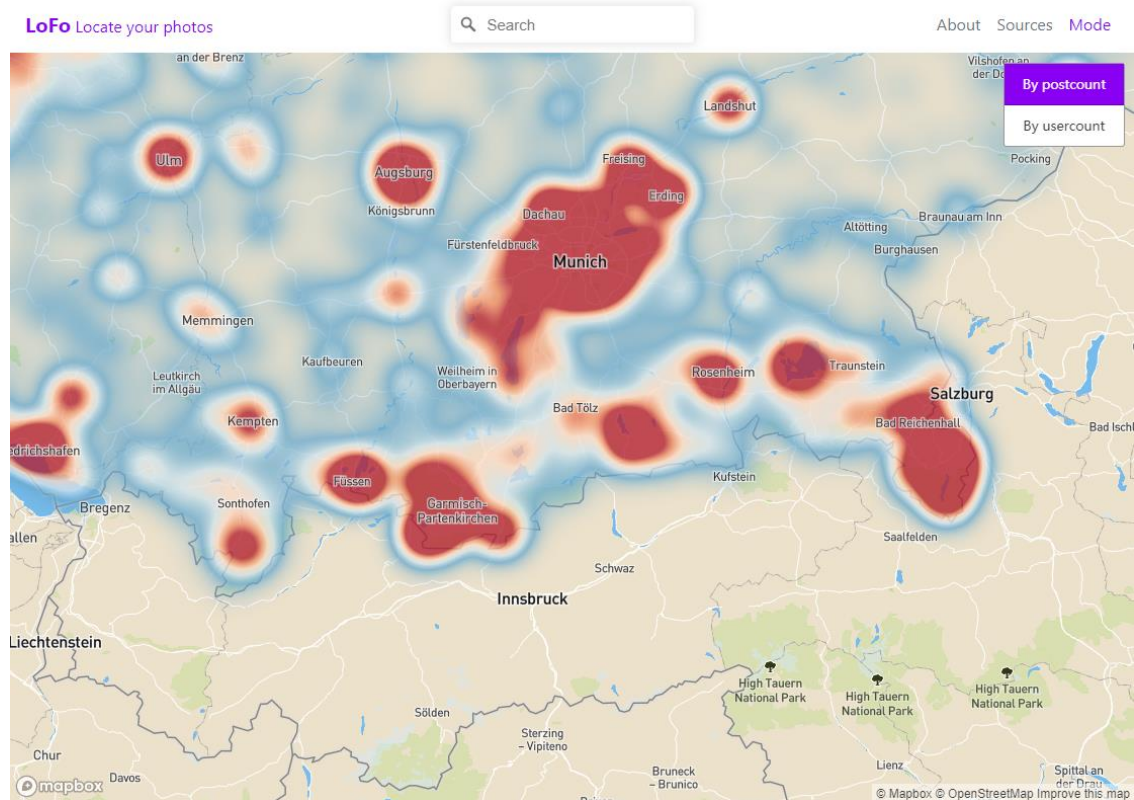


Figure 13 AOIs around Füssen and Garmisch-Partenkirchen using PC metric

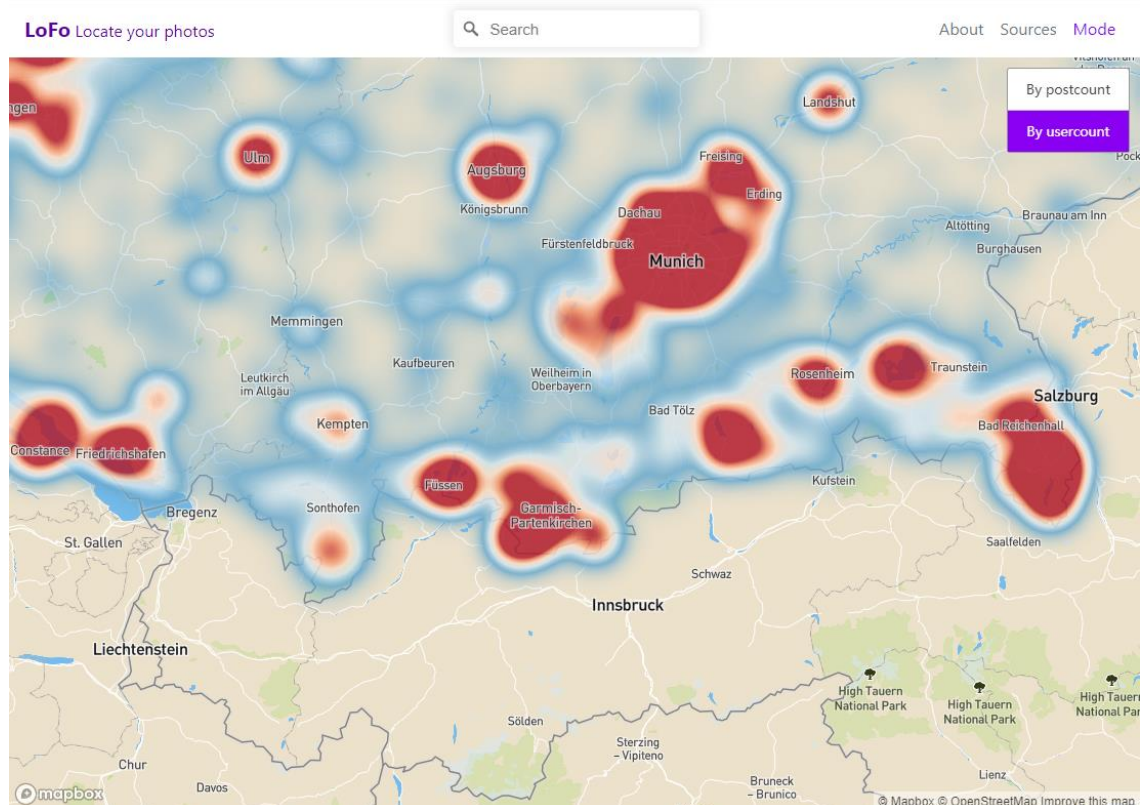


Figure 14 AOIs around Füssen and Garmisch-Partenkirchen using UC metric

4.2.2 AOI Choropleth Map

In order to give the users a clear perception of the geographic locations, choropleth maps are also used to combine AOIs with existing, known administrative boundaries. Using NUTS standard, it is easy for users to associate an area to a specific federal state, a government region, a district, or a postcode area under different NUTS levels. Figure 15 displays the initial state of the choropleth map mode, which offers the audience a summary of Germany at zoom level 4. Hovering on this area, information including population together with posts and users in total and per capita, is shown on the information window in the left top corner of the map display area.

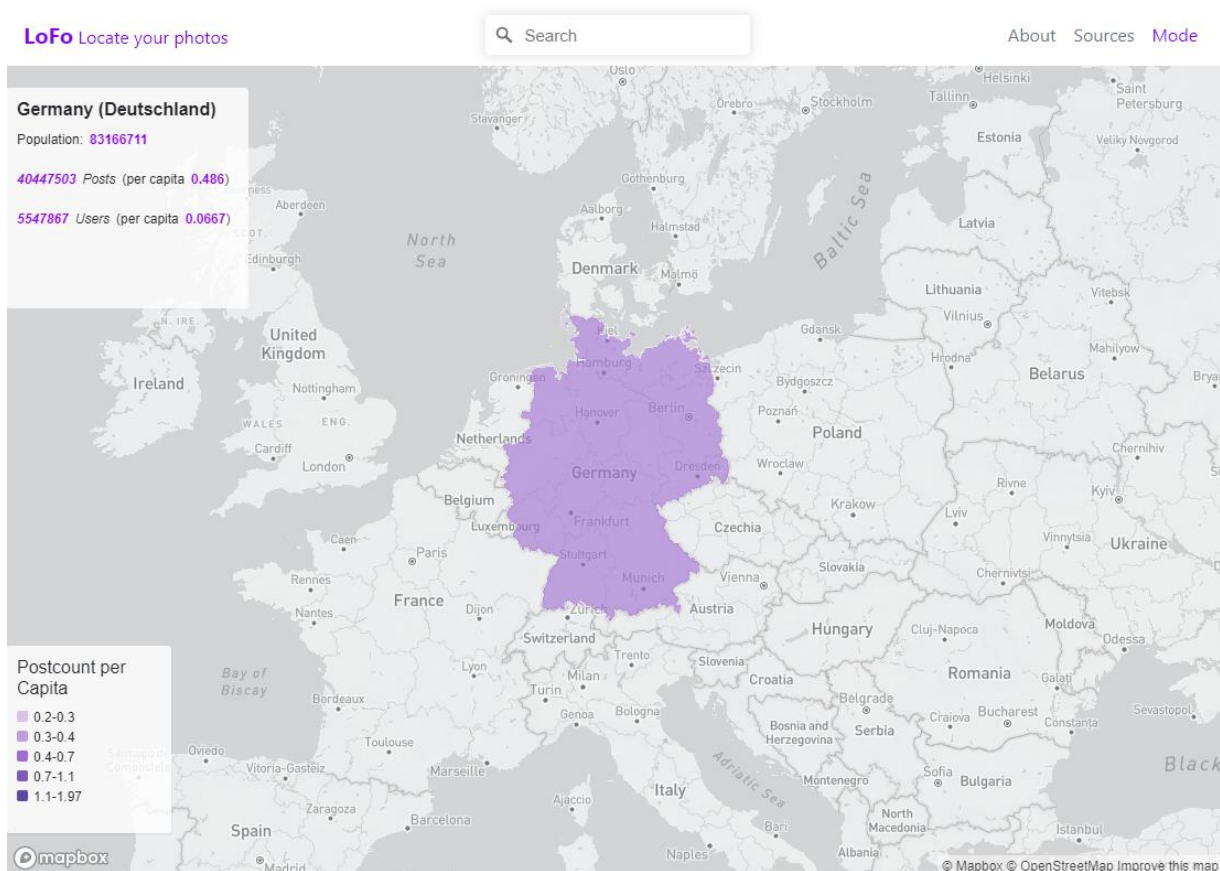


Figure 15 The initial state of the choropleth map showing a summary of Germany

Figure 16 demonstrates the effect after zooming in to aggregation on NUTS1 level, which provides an overview of 16 federal states of Germany. This map layer is visible between zoom level 5 and zoom level 6. This choropleth map uses a gradient color ramp including five different colors for symbolization based on postcount per capita instead of absolute values of postcount to balance the huge population and land area differences among states, while usercount per capita is also applicable. The map legend demonstrating colors and corresponding postcount per capita values is settled in the bottom left corner of the map display area. In this map, as the smallest federal state in Germany

apart from city-states, Saarland has the lowest postcount per capita value at 0.203. On the contrary, federal states such as Hamburg, Berlin, and Saxony can be easily distinguished as AOIs by their darker colors due to their relatively higher postcount per capita values.

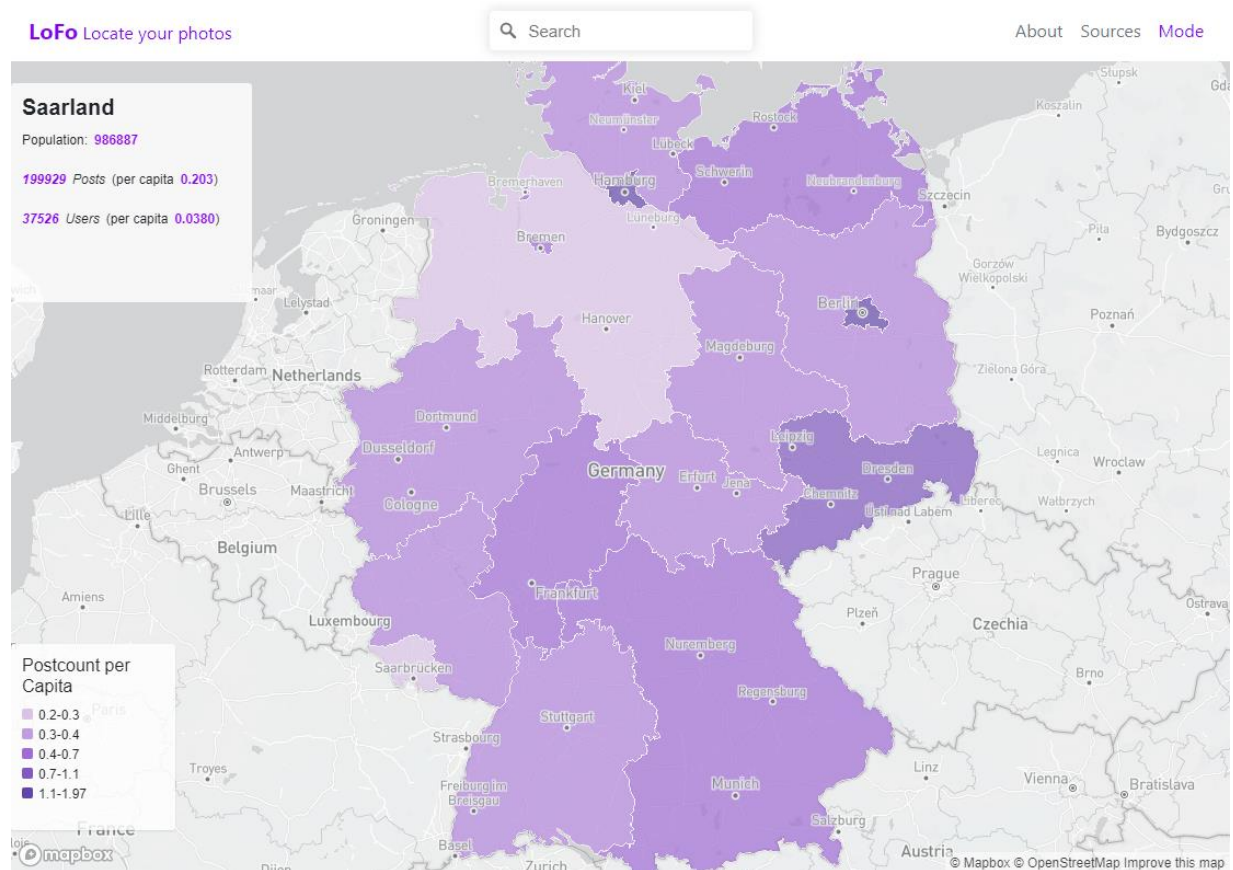


Figure 16 Choropleth map with an aggregation on NUTS1 level

The map layer in figure 17, which is visible between zoom level 6 to zoom level 7, is the visualization of data aggregated based on NUTS2 level containing 38 government regions (German: Regierungsbezirk) or administrative regions (German: Direktionsbezirke). Besides city-states, although several federal states cannot be split into government regions or administrative regions, like Brandenburg, Mecklenburg-Vorpommern, and Schleswig-Holstein, the introduction of this level does allow for a more fluid and consistent change in zoom levels. Since some of them are named as cities within the regions such as Dresden, Cologne, and Stuttgart, users can also easily associate the AOIs on this level to adjacent areas around some famous cities.

Introducing this layer not only allows for a smoother transition between the NUTS1- and NUTS3-level visualization but may also be able to serve visitors who want to accommodate in a city and take multiple day trips to explore the surrounding area. For exam-

ple, tourists could stay in Dresden and take day trips to areas such as Meißen, Saxony Switzerland national park, etc., which are less than three hours one way by train or car.

Same as the last map layer, map legend is located in the bottom left corner of the map display area. At the same time, the information window in the top left corner of the map display area displaying information that varies along users' mouse hovering area (users' clicking area on a touchscreen). Overall, Detmold, symbolized with white color, has the lowest postcount per capita value at 0.189 within the 38 regions in Germany. Regions including Leipzig, Dresden, and city-states like Hamburg and Berlin are symbolized with darker colors due to their high postcount per capita values that are all over 1. In the midst of these regions, Berlin has the highest postcount per capita value at 1.972.

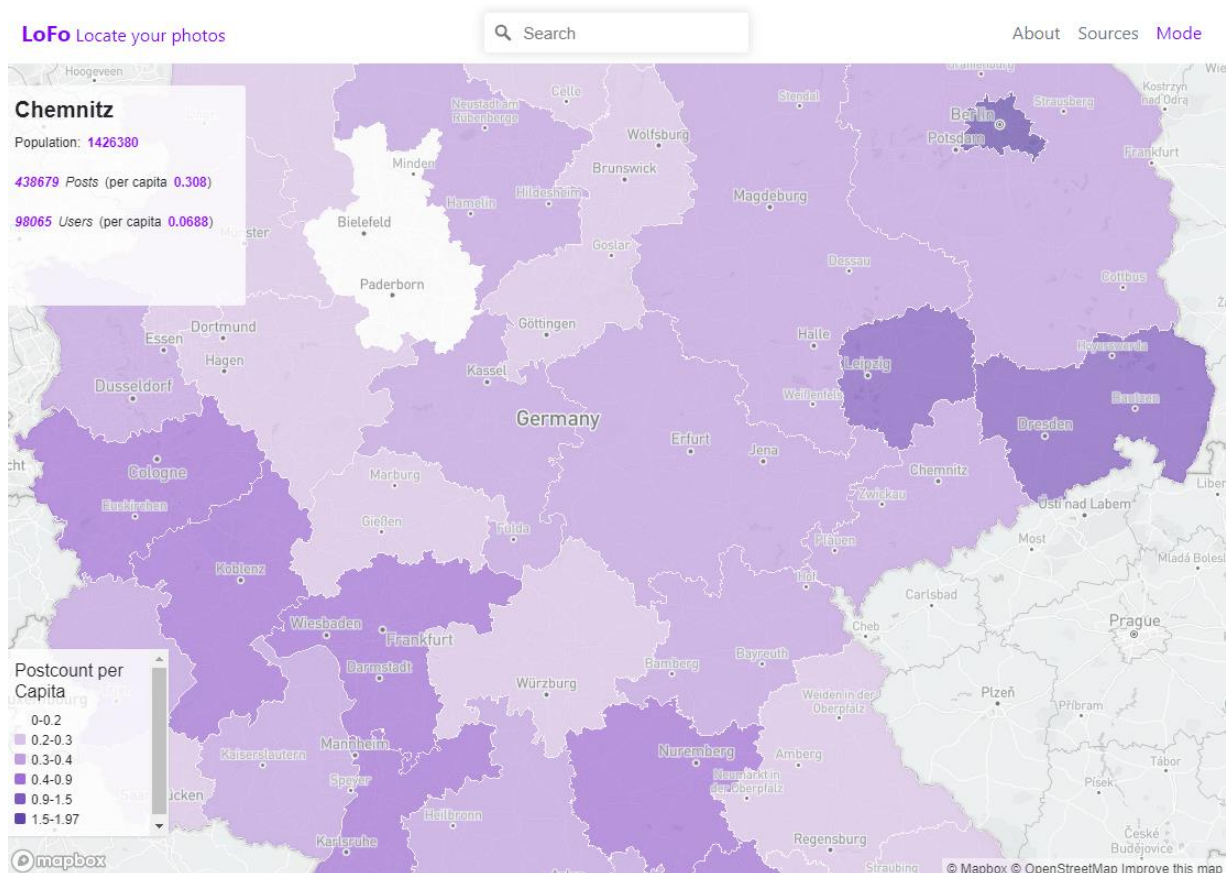


Figure 17 Choropleth map with an aggregation on NUTS2 level

The map layer in figure 18 is visible between zoom level 7 to zoom level 8. There are 401 districts in total within Germany when visualizing from data aggregated on NUTS3 level. These districts are known as Kreise (English: districts) or as kreisfreie Städte (English: independent cities) in German and generally have a population of 150,000 to 800,000 inhabitants (Destatis, n.d.). Zooming into southeast Germany, figure 18 illustrates similar results as in figure 13 and figure 14. AOIs such as Ostallgäu, which contains the area around Füssen, Garmisch-Partenkirchen, Munich (independent city), and Berch-

tesgadener Land, have postcount per capita values over 1.4 and are marked with darker purple colors. Within this range, districts such as Rottal-Inn, Augsburg (district), and Landshut (district) have relatively low postcount per capita values below 0.1.

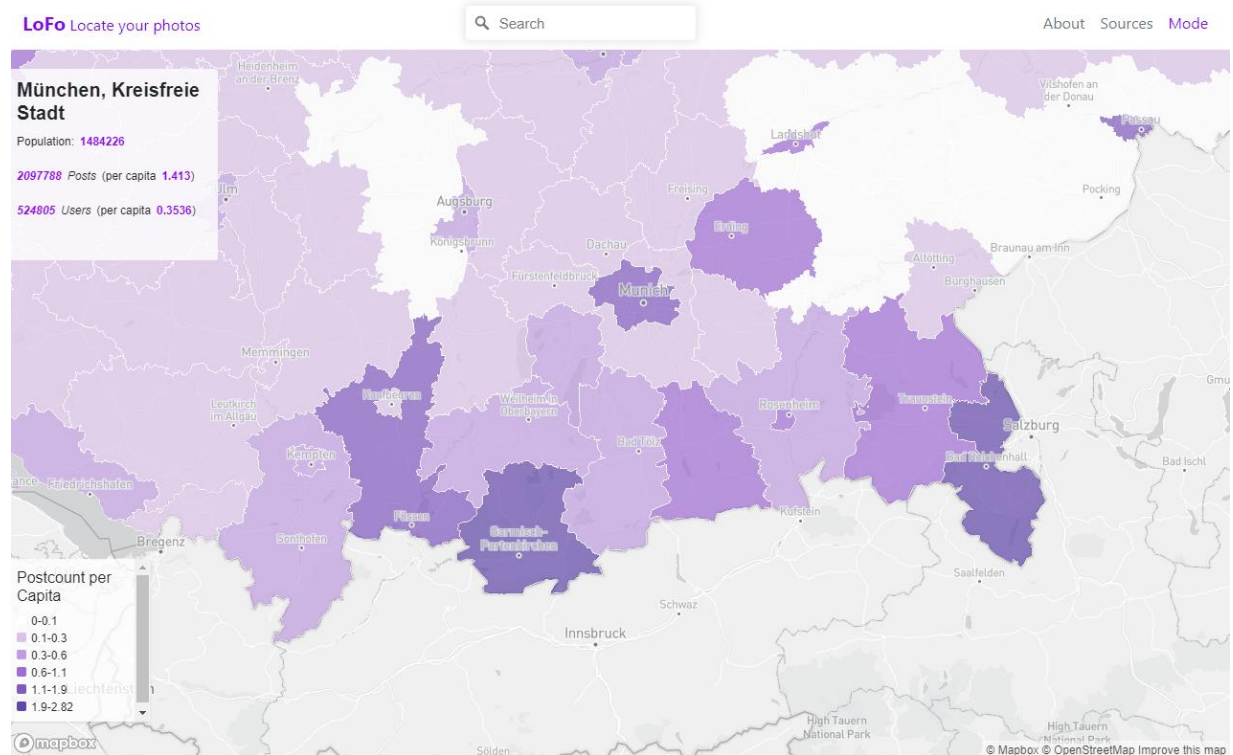


Figure 18 Choropleth map with an aggregation on NUTS3 level

Figure 19 is the visualization of areas around Dresden based on data aggregated on postcode. Compared with municipalities (German: Gemeinden) in Germany, postcodes are applied more frequently in real-life scenarios such as during delivery and filling in forms. On this level, since population and land area are comparably more balanced than the three NUTS levels, postcount instead of postcount per capita are used for symbolization. Besides, there are postcode areas that have zero or extremely few inhabitants, such as 30669 Langenhagen (Flughafen) and 82475 Garmisch-Partenkirchen (Schneefernerhaus), which would make the postcount per capita illogically high, causing misleading judgment. In this map layer, the information window and map legend interacting with users are located at the same positions. Although only a postcode and a city or district name are shown in the information window implying the area, users can still be informed about the geographic locations by the overlapping base map layer. Within Dresden, areas of which the postcode is 01067 and 01069 have more posts (322993 and 266342 separately) than other areas. This map layer may be able to help tourists choose their accommodation locations. Staying in a hotel located in an AOI in the city gives tourists more opportunities to explore POIs within walking distance.

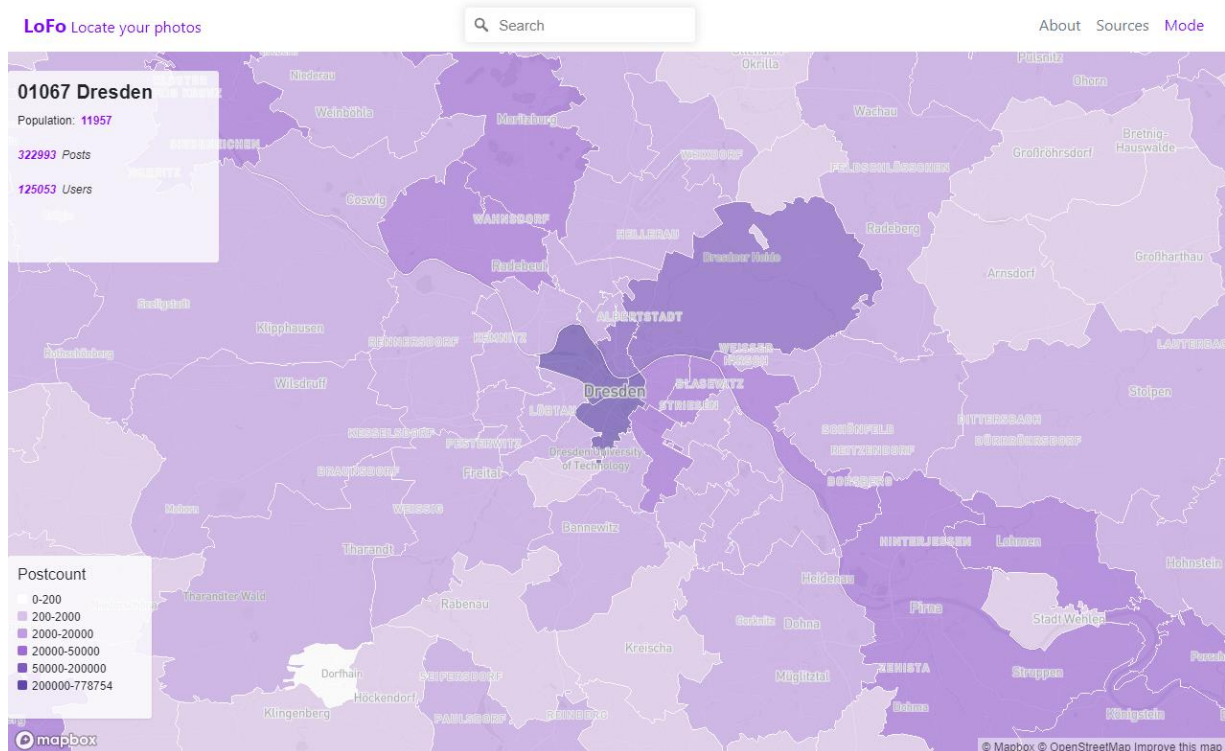


Figure 19 Choropleth map with an aggregation on postcode

4.3 Visualisation for Local POIs

4.3.1 Study case in Dresden

The study area of this case is Dresden, which is the capital city of the German federal state of Saxony. Having a long history, fabulous museums, and fantastic architecture, Dresden attracts over four million overnight stays every year. This map layer in figure 20 is visible at a minimum zoom level of 10. It visualizes all the local POIs based on the Flickr geotag data after density-based clustering and allows users to click on a POI to get more information from a pop-up window. These POIs mainly distribute in Innere Altstadt where historical architectures are situated, Neustadt where there are numerous restaurants and bars, and Großer Garten. In this map layer, the circle size of POIs varies along with zoom level and postcount, while the color of circles from white to dark red (color reference #B2182B) is only determined by postcount. On the same zoom level, the more posts that a POI has, the larger the circle is and the darker color the circle has.

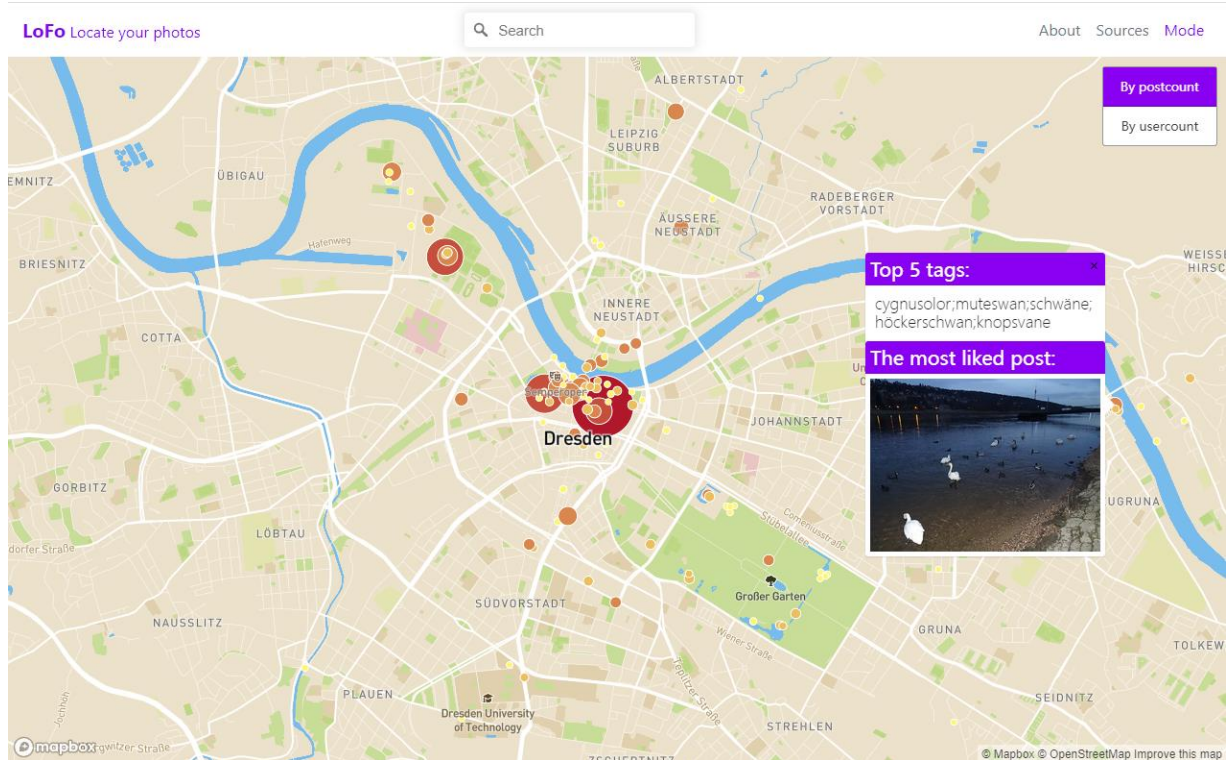


Figure 20 Local POIs in Dresden

On this level, the locations of POIs are relatively accurate. For example, when zooming in to the city center of Dresden, as in figure 21, the locations of POIs can be spotted clearly by overlapping with the base map. POIs located at Church of Our Lady (German: Frauenkirchen) and Zwinger Palace are relatively noticeable due to their comparably large circle size and dark circle color. Other POIs such as Brühl's Terrace, Semperoper, and Katholische Hofkirche can also be easily distinguished. When clicking on a POI circle, a pop-up window appears and displays the most frequently used five tags in Flickr posts and a thumbnail of the photo from the post which collected the most likes. The photo's thumbnail is also clickable, which sends the users to the original post on Flickr when the users are interested in this photo and would like to know more details. But all the images and original posts are only displayed when the viewing privacy set by the authors is public. This complements more information for POIs, provides more possibilities for interaction, and allows users to have a deeper insight into local POIs. The purple theme color is also applied here to highlight the two parts of this pop-up window.

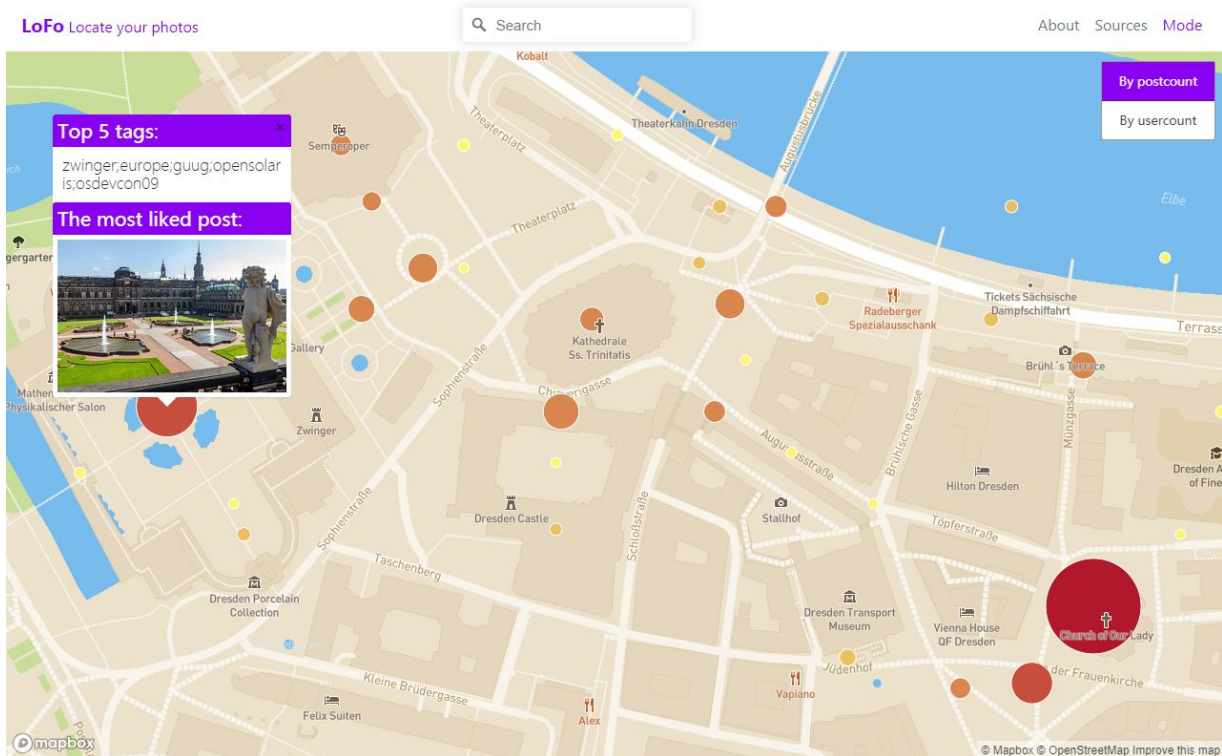


Figure 21 Local POIs in Dresden city center

4.3.2 Study case in Berlin

The other study case of local POI visualization is within Berlin city. As the most populous city in the European Union and a world city of history, culture, and politics, total tourist arrivals reached 14 million in the year 2019. This map layer, as shown in figure 22, is similar to the last map layer, which is visible at a minimum zoom level of 10. It visualizes all the local POIs based on the aggregated place dataset, and users can check the place name and postcount number of a POI by clicking it. Compared with the study case in Dresden, since this visualization is based on a much larger volume of data and because of the popularity of Berlin, more POIs can be viewed within Berlin city. Similarly, the circle size of POIs varies along with zoom level and postcount, and only postcount decides the color of circles. Famous tourist attractions like Checkpoint Charlie, Potsdamer Platz, Memorial to the Murdered Jews of Europe, etc., which attracted millions of posts, are all prominent POIs in this map layer.

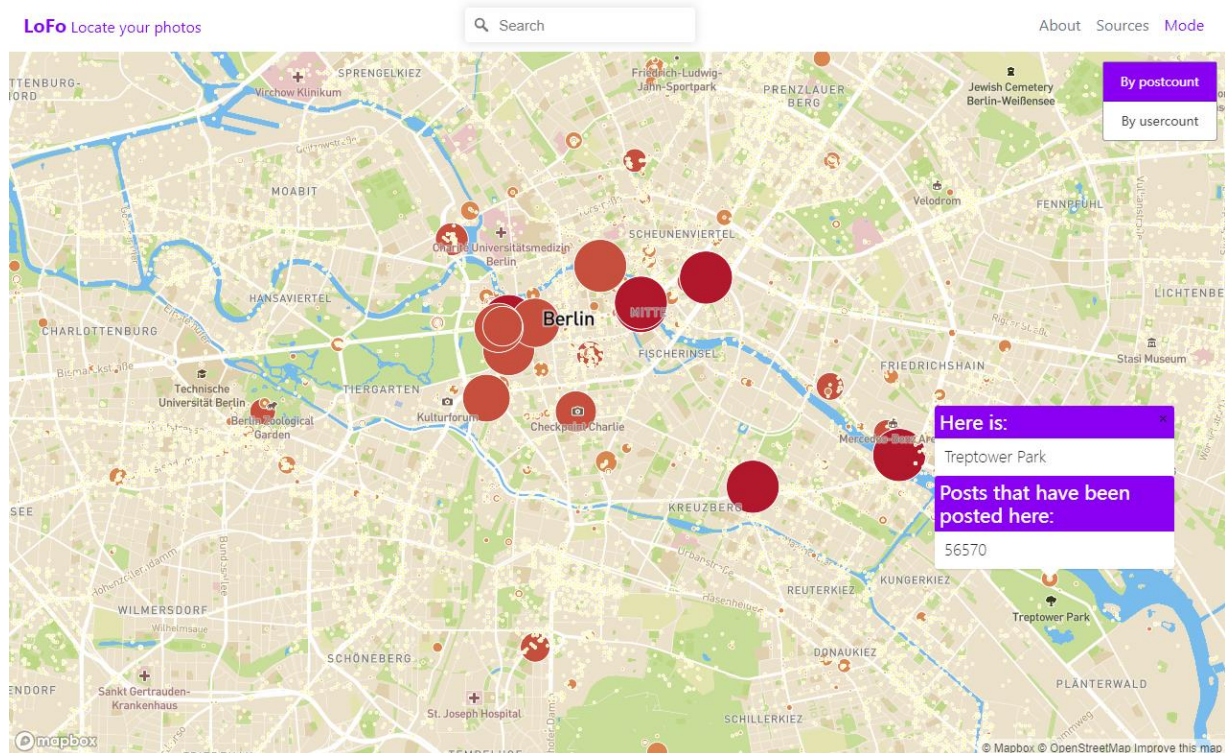


Figure 22 Local POIs in Berlin

Since the postcount value varies over a considerable range and there are quite a few locations with postcount values smaller 1000, these locations are visualized as small light yellow dots on the map. This way they become less visible compared to the more popular locations and do not obscure the more popular locations. Also, the information about these less popular locations can be preserved on the map without making the user feel overwhelmed by too much information.

On this level, locations of POIs are also accurate, with a location accuracy of approximately 19 meters. For instance, when zooming in to the area around Brandenburg Gate, the local POIs such as Berlin Wall, Brandenburg Gate can be spotted easily due to their noticeable circle size and dark circle color. As one of Berlin's most famous landmarks, it attracts numerous tourists and photographers to visit, take photos, and publish posts there. As shown in figure 23, there are over ten POIs around Brandenburg Gate with the same or similar places names, including Brandenburger Tor, Berlin Brandenburger Tor, Brandenburg Gate, etc. Tourists and photographers shoot Brandenburg Gate from various spots and different angles, which actually provides users more possibilities to explore multiple great photo shooting spots around specific landmarks. When clicking on a POI circle, a pop-up window appears and displays the place name and the number of posts that have been posted geo-tagging this place. Embedding these place names and postcount numbers in pop-up windows avoids overwhelming users with too much information at once and creates more interactions with users depending on their needs.

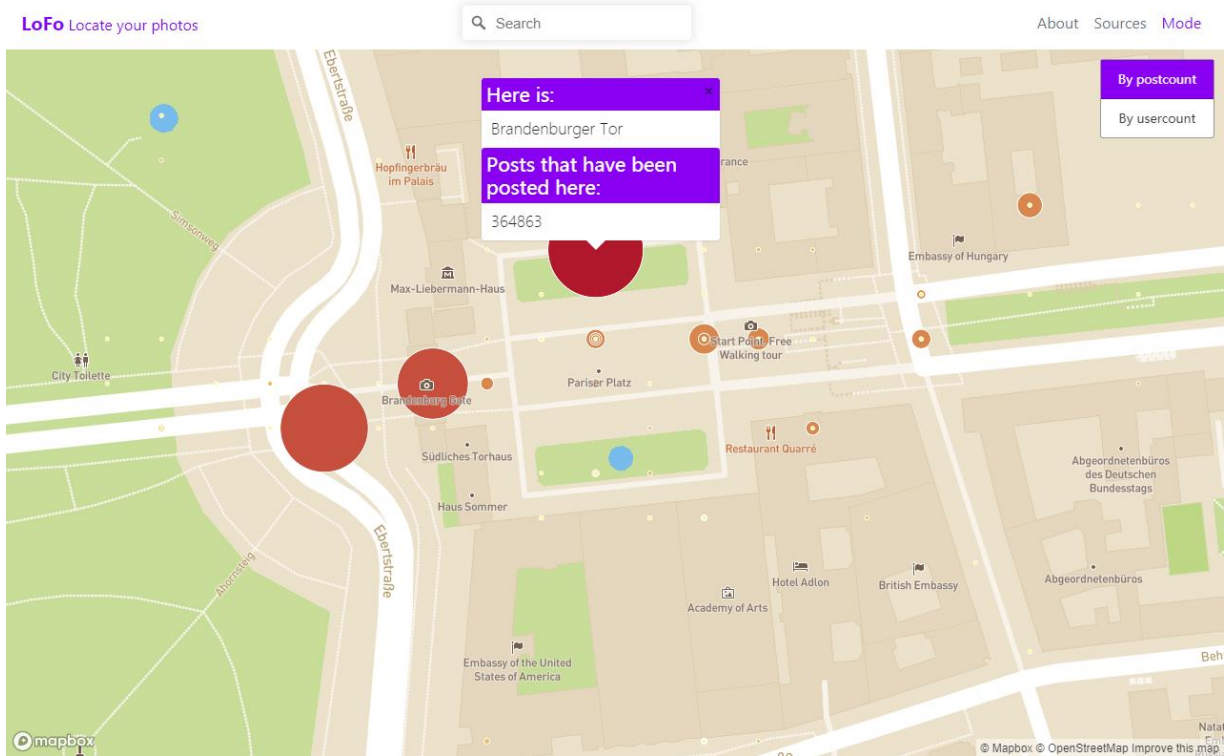


Figure 23 Local POIs around Brandenburger Tor

5 Discussion

5.1 Answers to the Research Questions

This research managed to achieve the research objective set in section 1.2, and eight research questions are discussed and answered during this study. The following is a summary of the answers to the research questions.

I . Identify the needs of visualizing POIs or AOIs for tourists and describe the data:

(a) For what purposes are tourists using visualizations of POI or AOI, and what are the requirements on different map scales?

Tourists use the visualization of AOIs or POIs during the initial planning phase of a trip and when selecting the locations to visit during an itinerary or upon arrival in the region. On smaller-scale maps, the visualization results should be presented as AOIs, i.e., visitors can explore areas of interest to them, corresponding to Germany, which can be federal states, government regions, or districts. On larger scale maps, such as in web map applications, when the zoom level is greater than 10, and major streets are visible, the visualization results should be presented as POIs. Users should be able to view specific locations within the city, such as highly visited churches, shopping malls, fairs, etc. Additionally, these popular locations should be supplemented with information as needed.

(b) What are the pros and cons of combining data from multiple social media platforms for multi-scale extraction and visualization of POIs for tourists?

Combining geo-tagged data from different social platforms can help encompass various user groups. It can also increase the data volume to a certain extent when the percentage of geo-tagged data is not high compared to the total data volume. However, for example, according to the product characteristics of Twitter, some of the tweets from Twitter may not be related to the geo-tagged location or not for travel purposes, and this kind of data is difficult to filter.

(c) How is the data structured, and what is the volume of available data?

The details about data structure and data volume are introduced in subsection 3.1. The number of posts included in each of the three VGI datasets is above a million. Additionally, the VGI data has already been preprocessed as the privacy-aware data structure.

(d) What parts of the data are related to either objective or subjective information?

Subsection 3.1 explains the definition of objective and subjective information: “if the data field reflects physically quantifiable properties of the environment or hardly quantifiable qualities.” And data fields are categorized according to that.

II. Select approaches of summarizing and aggregating POIs or AOIs from VGI:

(e) What is the difference between different metrics, e.g., User Count, Post Count, User Days?

Under the context of this research, the “Post Count” of each record in the dataset means the number of posts created and geotagged with the corresponding locations. Accordingly, “User Count” represents the number of users that have created a post or posts at the location. “User Days,” which is not utilized in this study, is the cumulation number of distinct user count each day at the site.

(f) What methods or algorithms should be employed while summarizing POIs or AOIs for different map scales?

The methods were selected based on the principle that they could be integrated into an easy-to-use workflow that is also able to achieve relatively good results. For summarising AOIs on smaller-scale maps, grid-based aggregation and administrative boundaries-based aggregation are applied separately to the aggregated post dataset. When the postcount value of each place is available, the dataset can be directly visualized as POIs on larger-scale maps, with the postcount values providing the popularity information. Additionally, if all the posts are available, a density-based clustering with an appropriate selection of parameters can be applied to aggregate the posts to visualize POIs on local levels.

III. Create the interactive visualization for the POIs and AOIs:

(g) How can POIs and AOIs be visualized on maps on different scales? Which information is important on which scale?

Data after grid-based aggregation or administrative boundaries-based aggregation are visualized and produced as two types of maps, which are heatmap and choropleth map, to summarise AOIs on a smaller scale. Under this context, the approximate extent of the areas, indicators of the popularity of the areas such as color, and at least one known place name within one AOI to indicate the location are required. Data after density-based clustering are visualized as proportional symbol maps on larger scales on which the locations and names of specific places on the map, the popularity of the locations characterized by the size and color of the circles, and additional information about the associated POIs are all critical.

(h) Are there necessary map elements and map interactive actions that can be included while visualizing POIs for tourism purposes, and how will they need to be implemented in real scenarios?

The web map application includes, in addition to the interactive map, a map legend, information display window, mode and layer toggle buttons, and a search box. Possible interactions include zoom in and out, pan, mouse hover, mouse click, and type search. Users can use these interactions to explore POIs under multi-scale maps, summon hover and pop-up windows to get more information about places and search to jump to known POIs and view the surrounding area.

5.2 Evaluation of the privacy protection

For grid-based aggregation and administrative boundaries-based aggregation, using HLL set union operation enables efficient and lossless calculation of the number of distinct users and posts per region/grid. In terms of user privacy protection, the aggregated place name dataset does not contain any personal user information. Since the output is a map containing only popularity and geographic location information, the original posts of users cannot be traced, thus achieving good privacy protection effect. For the aggregated post dataset, since the HLL data structure is applied and the original dataset has been aggregated, the data analyst is isolated from the raw data containing the user's personal information such as user id, post id, and any data analysis and visualization based on this dataset will not expose the users' personal information.

However, the user ids and post ids are included in the Flickr CCBY post dataset, which makes it tricky to protect the privacy of users in this dataset. After density-based clustering, three types of information are publicly visible in the output map: the popularity of POIs, the five most frequently used tags for each POI, and the images with the most likes for each POI. Although the images used are set to be publicly visible by the author, it is not allowed to use these images for any commercial purpose without the author's permission. In the context of this study, the use of these images is tolerable but still violates the principle of privacy protection. A better solution for POI information enhancement is to use images from Wikipedia, where the "public domain images are not copyrighted, and copyright law does not restrict their use in any way" (Wikipedia, n.d.). Since the filtered top five tags usually contain the place name information of the location marked, for POIs with high visitation rates, the place name information can be linked to the corresponding Wikipedia entry so that users can get more information and images about the location.

5.3 Evaluation of Data Aggregation and POI extraction

Aggregation based on two grid sizes of 20 km and 4 km allows the results to be visualized on maps of different scales. Grid-based aggregation using Geohash is highly easy to use, and by creating `geohash_reduce` function, this type of aggregation is straightforward to implement. However, due to the accuracy character of Geohash, this type of aggregation is limited and not flexible enough to customize the grid sizes according to the area of the study region, and only the grid sizes in Table 7 are available. Using the `width_bucket` function provided by PostgreSQL allows flexible partitioning of the study area according to needs. However, it is difficult to apply this method to make the segmented mesh equal in length and width, and the irregular shape may have an impact on the subsequent visualization.

Another possible solution is to create multiple shapefiles that contain the required different sizes of grids in GIS software such as ArcGIS Pro and QGIS, and then use them for grid-based aggregation. This method is more time-consuming but relatively accurate and flexible.

The accuracy of aggregation based on administrative boundaries is somewhat influenced by the data points located close to the boundaries. Since the data set used has already been aggregated, for example, some points located in area A may contain posts and users in A's neighborhood B, but this effect is fairly small for data aggregation above the districts.

From the results, both aggregations successfully highlight known tourist cities and or densely populated areas in Germany, such as Berlin, Hamburg, Munich, etc. Nonetheless, no specific popularity criteria for AOI were defined in this study for the results obtained from these two aggregations. So rather than extracting AOIs, the results of data aggregation and visualization give users the possibility to explore and compare different areas according to their own criteria and gain the AOIs in this way.

Since density-based clustering requires an understanding of the data and the selection of appropriate parameters, a corresponding criterion for "defining" a POI within the Dresden range is generated, i.e., a point density greater than or equal to thirty within a ten-meter radius of the post marker is extracted as a POI. The results obtained with this set of parameters are relatively satisfactory according to table 10, for example, when compared with Figure 24, which shows the tourist attractions of Dresden in Innere Altstadt, the POIs extracted by density-based clustering (see Figure 21) are highly compatible with it.

However, the criterion suited for Dresden city is not necessarily applicable to all the cities and rural regions in Germany, and the selection of parameters needs to be adapted

to the local context. For example, in sparsely populated areas, ϵ (eps) should be increased, and minPts should be decreased in order to extract relatively more meaningful results.

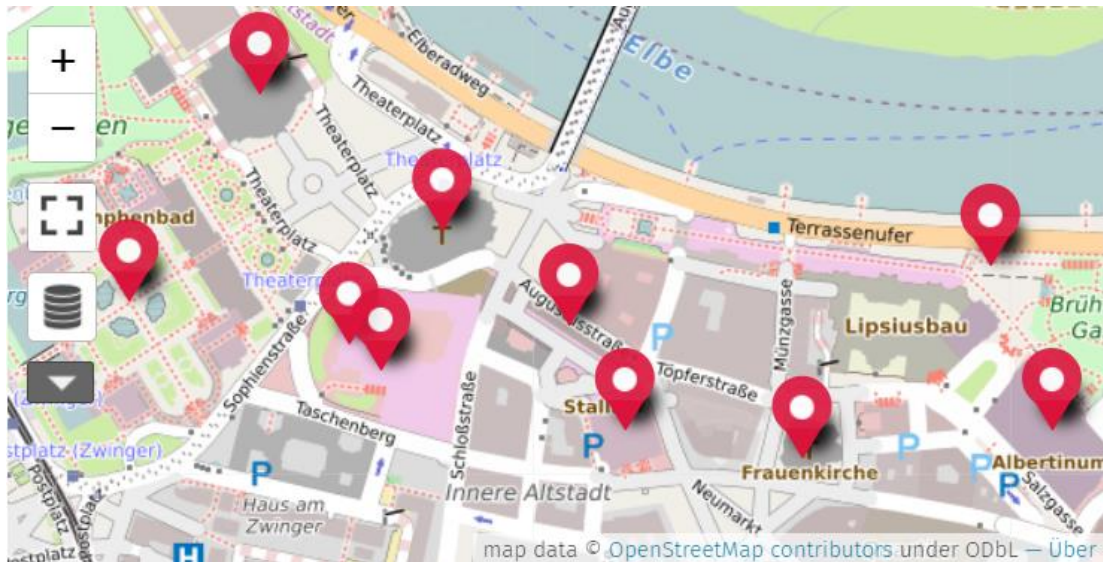


Figure 24 Tourist attractions in Dresden Innere Altstadt

5.4 Limitations of the Web Map

Legend is one of the essential map elements, but in the heatmap in this web map application, the legend is not included. Since this is a multi-scale map that can be zoomed in and out, the regions on the hotspot map are colored with a gradient effect to connect the various scales smoothly. This makes it challenging to create map legends for such applications compared to static maps. Nevertheless, in this customized application for tourists, a uniform map legend is not necessary to help quantify the AOIs, as there is no uniform standard for extracting them, and the exploration and perception of latent information for this application is obtained by comparison with neighboring areas.

In choropleth maps with aggregation based on administrative boundaries, the map legend is created and placed in the lower-left corner of the map. Although the maps displayed on this page are still multi-scale, the lack of gradients between maps as they are zoomed in and zoomed out makes it possible to create map legends. Due to the different criteria for color selection between different scales, for example, at the postcode level, the coloring is based on postcount. In contrast, the other higher levels are based on postcount per capita, as in figure 25. The combination of the map legend and the information window in the upper left corner allows the user to clarify the meaning of the colors and alleviate any misunderstandings caused by the change in metrics.

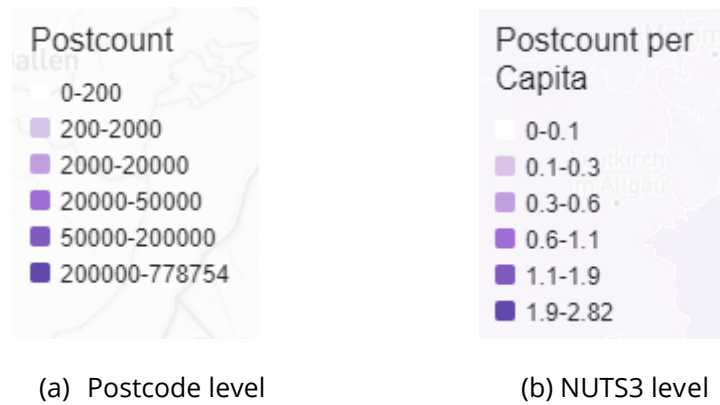


Figure 25 Map legends in choropleth maps

In addition, the user guide has been placed on the “About” page to make the interface more concise. This also makes it confusing and takes some time for users who have not used web mapping applications to explore the application for the first time. To compensate for this, the application uses a purple theme color to highlight areas that need to be clicked on to guide the users.

6 Conclusion and Outlook

6.1 Conclusion

This research developed a workflow to visualize and summarize POIs for tourist guiding purposes, on a multi-scale and national-range map, based on three national VGI datasets: aggregated place name dataset, aggregated post dataset, and Flickr CCBY post dataset. Grid-based aggregation, administrative boundaries-based aggregation, density-based clustering are applied separately to aggregate the VGI data on multiple map scales. This study created an interactive web map application as an output, which targets to serve tourists and photographers during the planning phase of a trip or a photoshoot. The application has a straightforward interface design, and thanks to its interactive nature, users can access more information than the current map display area through various interactive methods such as mouse hover, mouse click, zoom in and out, and keyboard typing and searching. By combining three different types of multi-scale maps, the application gives users a good overview of POIs within Germany. Furthermore, it allows users to explore and compare POI extraction results using different metrics for visualization at different scales.

6.2 Future work

In general, the approach adopted in this study allows for the summary and extraction of POIs in Germany. However, the density-based clustering approach was only experimented in Dresden and not validated in other cities or rural areas. And since the grid-based aggregation uses only two sizes of the grid, adding a few layers of grid-based aggregation between or beyond the two sizes might make the transition between scales more smooth. In addition to that, the current NUTS 2021 classification is in effect as of January 1, 2021 (Eurostat, n.d.). Since NUTS standard lists 92 NUTS1 level areas, 242 NUTS2 level areas, and 1166 regions at NUTS3 level, which cover the territory of the EU members and the UK, theoretically, administrative (NUTS) boundaries-based aggregation can also be applied to EU countries and the UK to analyze and visualize VGI data.

Based on that, the possible recommendations for future work can be:

- Validate the feasibility of density-based VGI data clustering for POI extraction in selected cities and rural areas within Germany and evaluate the quality of the results.
- Use GIS software to generate shapefiles with different size grids in Germany and aggregate and visualize VGI data on this basis.

- Obtain VGI data for other EU countries and the UK and explore the feasibility of NUTS boundaries-based aggregations in these regions.
- Improve the user interface design of the web map application, make the guidance for users clearer and enhance user interaction.

References

- ArcGIS Desktop. (n.d.). What is an SRID. <https://desktop.arcgis.com/en/arcmap/10.3/manage-data/using-sql-with-gdbs/what-is-an-srid.htm>
- Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, 49, 45–53. <https://doi.org/10.1016/j.apgeog.2013.09.012>
- Bartoschek, T., & Keßler, C. (2013). VGI in Education: From K-12 to Graduate Studies. In D. Sui, S. Elwood, & M. Goodchild (Eds.), *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* (pp. 341–360). Springer Netherlands. https://doi.org/10.1007/978-94-007-4587-2_19
- Citusdata/postgresql-hll. (n.d.). Postgresql-hll extension. <https://github.com/citusdata/postgresql-hll>
- Corradi, A., Curatola, G., Foschini, L., Ianniello, R., & Rolt, C. R. D. (2015). Automatic extraction of POIs in smart cities: Big data processing in ParticipAct. 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), 1059–1064. <https://doi.org/10.1109/INM.2015.7140433>
- Crockett, K. A., Mclean, D., Latham, A., & Alnajran, N. (2017). Cluster Analysis of Twitter Data: A Review of Algorithms. *Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, 2, 239–249. <http://www.icaart.org/>
- Destatis. (n.d.). NUTS classification. https://www.destatis.de/Europa/EN/Methods/Classifications/OverviewClassification_NUTS.html
- Dunkel, A., Andrienko, G., Andrienko, N., Burghardt, D., Hauthal, E., & Purves, R. (2019). A conceptual framework for studying collective reactions to events in location-based social media. *International Journal of Geographical Information Science*, 33(4), 780–804. <https://doi.org/10.1080/13658816.2018.1546390>
- Dunkel, A., Löchner, M., & Burghardt, D. (2020). Privacy-Aware Visualization of Volunteered Geographic Information (VGI) to Analyze Spatial Activity: A Benchmark Implementation. *ISPRS International Journal of Geo-Information*, 9(10), 607. <https://doi.org/10.3390/ijgi9100607>
- Elwood, S. (2008). Volunteered geographic information: Future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal*, 72(3), 173–183. <https://doi.org/10.1007/s10708-008-9186-0>

- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).
- Eurostat (n.d.). NUTS Background. <https://ec.europa.eu/eurostat/web/nuts/background>
- Flajolet, P., Fusy, É., Gandouet, O., & Meunier, F. (2007). HyperLogLog: The analysis of a near-optimal cardinality estimation algorithm. In P. Jacquet (Ed.), *AofA: Analysis of Algorithms: Vol. DMTCS Proceedings vol. AH, 2007 Conference on Analysis of Algorithms (AofA 07)* (pp. 137–156). Discrete Mathematics and Theoretical Computer Science. <https://hal.inria.fr/hal-00406166>
- Flickr. (2020, November 23). Flickr Privacy Policy. <https://www.flickr.com/help/privacy>
- Flickr blog. (2011, August 3). Introducing geofences on Flickr. <https://blog.flickr.net/en/2011/08/30/introducing-geofences-on-flickr/>
- Flickr Services (n.d.). Flickr photo license information. <https://www.flickr.com/services/api/flickr.photos.licenses.getInfo.html>
- Gan, J., & Tao, Y. (2015). DBSCAN Revisited: Mis-Claim, Un-Fixability, and Approximation. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 519–530. <https://doi.org/10.1145/2723372.2737792>
- Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *Geojournal*, 69(4), 211–221. <https://doi.org/10.1007/s10708-007-9111-y>
- Granell, C., & Ostermann, F. O. (2016). Beyond data collection: Objectives and methods of research using VGI and geo-social media for disaster management. *Computers, Environment and Urban Systems*, 59, 231–243. <https://doi.org/10.1016/j.compenvurbsys.2016.01.006>
- Hauff, C. (2013). A study on the accuracy of Flickr's geotag data. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 1037–1040. <https://doi.org/10.1145/2484028.2484154>
- Hochmair, H. H., Juhász, L., & Cvetojevic, S. (2018). Data Quality of Points of Interest in Selected Mapping and Social Media Platforms. In P. Kiefer, H. Huang, N. Van de Weghe, & M. Raubal (Eds.), *Progress in Location Based Services 2018* (pp. 293–313). Springer International Publishing. https://doi.org/10.1007/978-3-319-71470-7_15
- Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, 54, 240–254. <https://doi.org/10.1016/j.compenvurbsys.2015.09.001>

- Huang, B., & Carley, K. M. (2019). A large-scale empirical study of geotagging behavior on Twitter. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 365–373. <https://doi.org/10.1145/3341161.3342870>
- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59–68. <https://doi.org/10.1016/j.bushor.2009.09.003>
- Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014). DBSCAN: Past, present and future. *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, 232–238. <https://doi.org/10.1109/ICADIWT.2014.6814687>
- Lerin, P. M., Yamamoto, D., & Takahashi, N. (2011). Inferring and Focusing Areas of Interest from GPS Traces. In K. Tanaka, P. Fröhlich, & K.-S. Kim (Eds.), *Web and Wireless Geographical Information Systems* (pp. 176–187). Springer. https://doi.org/10.1007/978-3-642-19173-2_14
- Liu, Y., Yuan, Y., Xiao, D., Zhang, Y., & Hu, J. (2010). A point-set-based approximation for areal objects: A case study of representing localities. *Computers, Environment and Urban Systems*, 34(1), 28–39. <https://doi.org/10.1016/j.compenvurbsys.2009.05.001>
- Liu, Y., Singleton, A., Arribas-bel, D., & Chen, M. (2021). Identifying and understanding road-constrained areas of interest (AOIs) through spatiotemporal taxi GPS data: A case study in New York City. *Computers, Environment and Urban Systems*, 86, 101592. <https://doi.org/10.1016/j.compenvurbsys.2020.101592>
- Lobo, M.-J., Pietriga, E., & Appert, C. (2015). An Evaluation of Interactive Map Comparison Techniques. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3573–3582. <https://doi.org/10.1145/2702123.2702130>
- Mapbox. (2021) Zoom levels in Mapbox. <https://docs.mapbox.com/help/glossary/zoom-level/>
- Meng, C., Cui, Y., He, Q., Su, L., & Gao, J. (2017). Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data. *2017 IEEE International Conference on Big Data (Big Data)*, 1319–1324. <https://doi.org/10.1109/BigData.2017.8258062>
- Moreri, K., Fairbairn, D., & James, P. (2018). Issues in developing a fit for purpose system for incorporating VGI in land administration in Botswana. *Land Use Policy*, 77, 402–411. <https://doi.org/10.1016/j.landusepol.2018.05.063>

- Popescu, A., & Shabou, A. (2013). Towards precise POI localization with social media. *Proceedings of the 21st ACM International Conference on Multimedia*, 573–576. <https://doi.org/10.1145/2502081.2502151>
- PostGIS. (n.d.). Spatial and Geographic Objects for PostgreSQL. <https://postgis.net/>
- R. (n.d.). The R Project for Statistical Computing. <https://www.r-project.org/>
- Reise magazin. (n.d.). Top 30 Dresden tourist attractions. <https://www.voucherwonderland.com/reisemagazin/sehenswuerdigkeiten-dresden/>
- Ricker, B. A., Johnson, P. A., & Sieber, R. E. (2013). Tourism and environmental change in Barbados: Gathering citizen perspectives with volunteered geographic information (VGI). *Journal of Sustainable Tourism*, 21(2), 212–228. <https://doi.org/10.1080/09669582.2012.699059>
- Roth, R. E. (2013). Interactive maps: What we know and what we need to know. *Journal of Spatial Information Science*, 2013(6), 59–115. <https://doi.org/10.5311/JOSIS.2013.6.105>
- Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems*, 42(3), 19:1-19:21. <https://doi.org/10.1145/3068335>
- Scikit-learn. (n.d.). Machine learning in Python—Scikit-learn 1.0.1 documentation. <https://scikit-learn.org/stable/>
- Smith, M., Szongott, C., Henne, B., & von Voigt, G. (2012). Big data privacy issues in public social media. 2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST), 1–6. <https://doi.org/10.1109/DEST.2012.6227909>
- Statista. (2021, May 20). German holiday trip organizing 2020. <https://www.statista.com/statistics/1031911/type-of-planning-organising-of-german-holiday-trips/>
- SUCHE-POSTLEITZAHL.ORG (2020, November 7). Postcode Germany. <https://www.suche-postleitzahl.org/plz-karte-erstellen>
- Sui, D. Z. (2008). The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems*, 32(1), 1–5. <https://doi.org/10.1016/j.compenvurbsys.2007.12.001>
- Sun, Y., Fan, H., Helbich, M., & Zipf, A. (2013). Analyzing Human Activities Through Volunteered Geographic Information: Using Flickr to Analyze Spatial and Temporal Pattern of Tourist Accommodation. In J. M. Krisp (Ed.), *Progress in Location-Based Services* (pp. 57–69). Springer. https://doi.org/10.1007/978-3-642-34203-5_4

- VGIscience. (n.d.). Basic visual analytics. <https://lbsn.vgiscience.org/yfcc-geohash/>
- Wikipedia. (n.d.). Image use policy. https://en.wikipedia.org/wiki/Wikipedia:Image_use_policy
- Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using social media to quantify nature-based tourism and recreation. *Scientific Reports*, 3(1), 2976. <https://doi.org/10.1038/srep02976>
- WorldAtlas. (2021, February 24). Map of Germany. <https://www.worldatlas.com/maps/germany>
- Yang, B., Zhang, Y., & Lu, F. (2014). Geometric-based approach for integrating VGI POIs and road networks. *International Journal of Geographical Information Science*, 28(1), 126–147. <https://doi.org/10.1080/13658816.2013.830728>
- Yang, W., & Ai, T. (2018). POI Information Enhancement Using Crowdsourcing Vehicle Trace Data and Social Media Data: A Case Study of Gas Station. *ISPRS International Journal of Geo-Information*, 7(5), 178. <https://doi.org/10.3390/ijgi7050178>
- Zheng, Y. (2011). Location-Based Social Networks: Users. In Y. Zheng & X. Zhou (Eds.), *Computing with Spatial Trajectories* (pp. 243–276). Springer. https://doi.org/10.1007/978-1-4614-1629-6_8

Appendices

Appendix A: SQL code for grid-based aggregation

To create geohash reduce (grid-based aggregation) function:

```
CREATE OR REPLACE FUNCTION
extensions.geohash_reduce (IN coord geometry, geohash integer DEFAULT 5)
RETURNS geometry
AS $$
    SELECT
        CASE WHEN ST_Y(coord) = 0 AND ST_X(coord) = 0
        THEN
            coord
        ELSE
            ST_PointFromGeoHash(ST_GeoHash(coord, geohash), geohash)
        END as "coord"
    $$
LANGUAGE SQL
STRICT;
```

To aggregate the data on approximately 20 kilometer-grids:

```
CREATE TABLE mdata.point_20km AS
WITH agg_20km AS (
    SELECT
        ST_Y(extensions.geohash_reduce(p1.the_geom, 4)) AS latitude,
        ST_X(extensions.geohash_reduce(p1.the_geom, 4)) AS longitude,
        userhll,
        posthll
    FROM data_original p1
)
SELECT
    agg_20km.latitude,
    agg_20km.longitude,
    hll_union_agg(agg_20km.userhll) AS user_hll,
    hll_union_agg(agg_20km.posthll) AS post_hll,
    hll_cardinality(hll_union_agg(agg_20km.userhll)) AS usercount,
    hll_cardinality(hll_union_agg(agg_20km.posthll)) AS postcount,
    ST_SetSRID(
        ST_MakePoint(
            agg_20km."longitude",
            agg_20km."latitude"),
        4326) AS latlng_geom
FROM agg_20km
GROUP BY agg_20km."latitude", agg_20km."longitude";
```


To aggregate the data on approximately 4 kilometer-grids:

```
CREATE TABLE mdata.point_4km AS
WITH agg_4km AS (
  SELECT
    ST_Y(extensions.geohash_reduce(p1.the_geom, 5)) AS "latitude",
    ST_X(extensions.geohash_reduce(p1.the_geom, 5)) AS "longitude",
    userhll,
    posthll
  FROM data_original AS p1
)
SELECT
  agg_4km."latitude",
  agg_4km."longitude",
  hll_union_agg(agg_4km.userhll) AS user_hll,
  hll_union_agg(agg_4km.posthll) AS post_hll,
  hll_cardinality(hll_union_agg(agg_4km.userhll)) AS usercount,
  hll_cardinality(hll_union_agg(agg_4km.posthll)) AS postcount,
  ST_SetSRID(
    ST_MakePoint(
      agg_4km."longitude",
      agg_4km."latitude"),
      4326) AS latlng_geom
FROM agg_4km
GROUP BY agg_4km."latitude", agg_4km."longitude";
```

Appendix B: SQL code for administrative boundaries-based aggregation

To aggregate the data on NUTS3 level:

```
CREATE TABLE data_landkreis AS
SELECT ld.nuts3 AS nuts3,
  ld.l_name AS l_name,
  ld.l_type AS l_type,
  ld.einwohner AS einwohner,
  ld.geom AS geom,
  sum(data_original.postcount) AS postcount,
  hll_cardinality(hll_union_agg(userhll)) AS usercount,
  CAST (sum(data_original.postcount) AS float)/ld.einwohner AS post_per_capita,
  hll_cardinality(hll_union_agg(userhll))/ld.einwohner AS user_per_capita
FROM data_original,
  public.c250_nuts3_project AS ld
WHERE ST_Within(data_original.the_geom, ld.geom)=TRUE
GROUP BY ld.nuts3,
  ld.l_name,
  ld.l_type,
  ld.einwohner,
  ld.geom
```

Appendix C: SQL code for tags ranking and filtering

```
/* connect the post tags of each cluster with ';' */
CREATE TABLE mdata.mtags AS
SELECT cid,
       string_agg(tags,
                  ';' )
FROM mdata.flickr10_30
WHERE cid IS NOT NULL
GROUP BY cid;

/* store each tag in a separate row */
CREATE MATERIALIZED VIEW mviews.mtags AS
SELECT *
FROM
  (SELECT cid,
          regexp_split_to_table(string_agg, ';') AS tag
   FROM mdata.mtags) t
WHERE t.tag IS NOT NULL
ORDER BY cid;

/* exclude the general tags */
CREATE MATERIALIZED VIEW mviews.mtags_fre1 AS
SELECT *
FROM mviews.mtags_frequency
WHERE tag not in ('dresden',
                  'germany',
                  'sachsen',
                  'saxony',
                  'deutschland',
                  '');

/* select the most frequently used five tags */
CREATE MATERIALIZED VIEW mviews.top5tags AS
SELECT cid, tag, tag_num
FROM
  (SELECT cid,
          tag,
          tag_num,
          ROW_NUMBER() over(PARTITION BY cid
                             ORDER BY tag_num DESC) AS ranks
   FROM mviews.mtags_fre1) AS a
WHERE a.ranks<=5
```