Cartography M.Sc.

Extraction of Places of Interest from VGI

Junru Lin

22.11.2021

Outline

- Introduction and Motivation
- Research Objective
- Research Questions
- Methodology and Application
  - Data
  - Data Aggregation
  - Interactive Visualization
- Discussion
- Conclusion and Outlook

Introduction and Motivation

Background:

- In many application scenarios, such as urban planning, traffic guidance, travel planning, POI (place of insterest) plays a vital role in supporting decision making.
- Following the rapid development and pervasive use of location-based technology, large volumes of spatiotemporal data from e.g. OSM and social networks offer new opportunities to visualize and understand urban dynamics and human movement (Arribas-Bel, 2014).

Introduction and Motivation

Research gap:

- Despite current trends in information visualization, POIs are still often displayed as pins on maps or as ranked lists of places for cities.

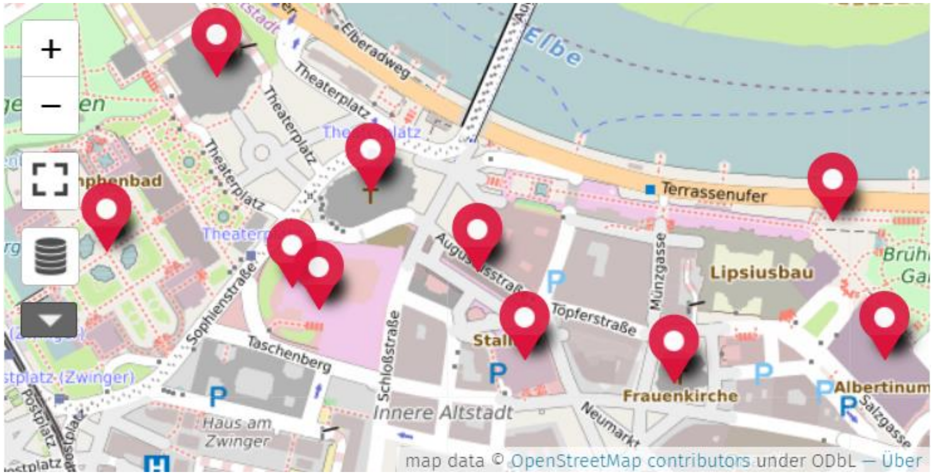


Fig. 1. Tourist attractions in Dresden Innere Altstadt

Introduction and Motivation

Research gap:

- 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.

Research Objective

- 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 three datasets derived from VGI (notably LBSM data).

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?

(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:

Research Questions

(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?

Methodology and Application

Workflow:

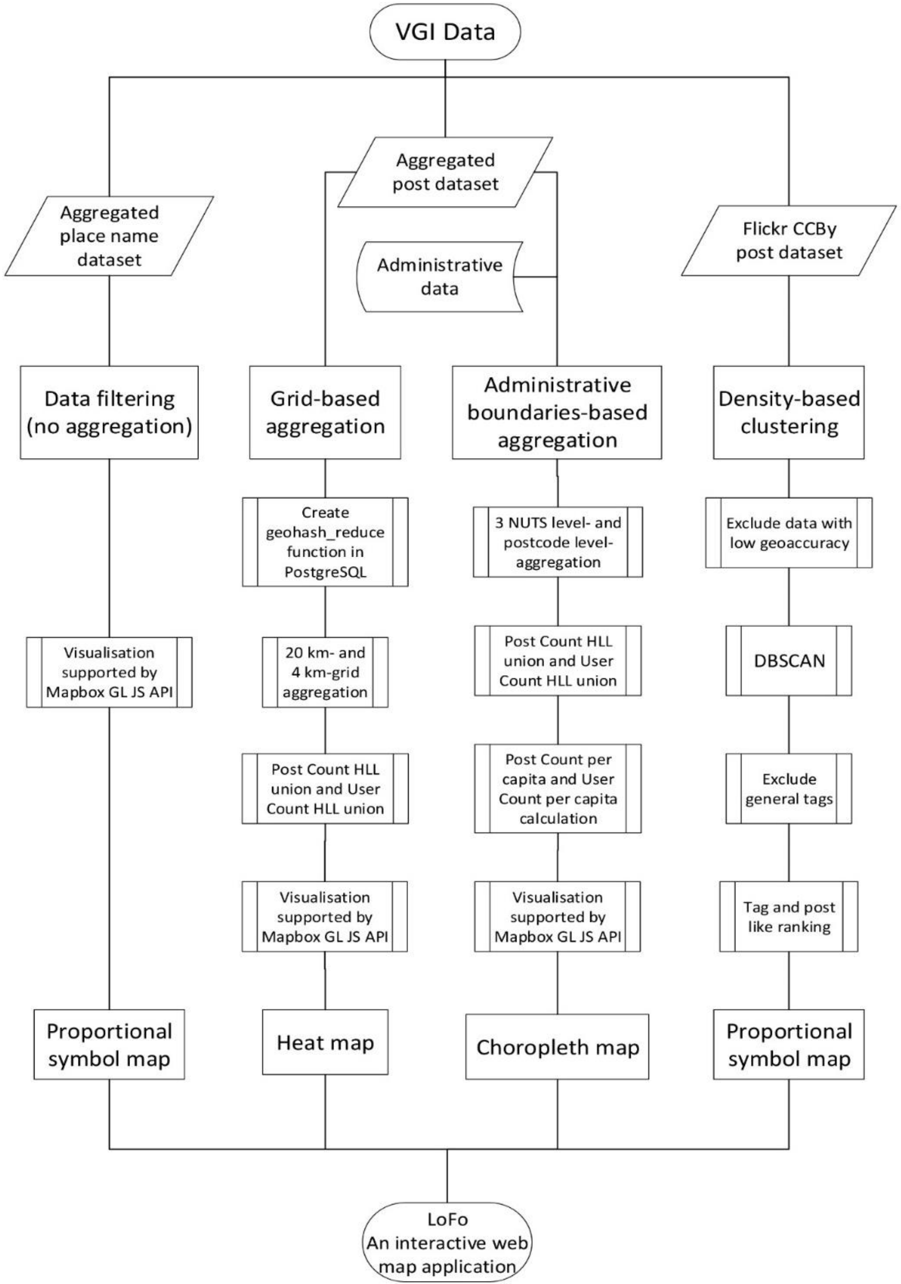

VGI Data
Aggregated place name dataset
Aggregated post dataset
Administrative data
Flickr CCBy post dataset
Data filtering (no aggregation)
Grid-based aggregation
Administrative boundaries-based aggregation
Density-based clustering
Create geohash_reduce function in PostgreSQL
3 NUTS level- and postcode level-aggregation
Exclude data with low geoaccuracy
Visualisation supported by Mapbox GL JS API
20 km- and 4 km-grid aggregation
Post Count HLL union and User Count HLL union
DBSCAN
Post Count HLL union and User Count HLL union
Post Count per capita and User Count per capita calculation
Exclude general tags
Visualisation supported by Mapbox GL JS API
Visualisation supported by Mapbox GL JS API
Tag and post like ranking
Proportional symbol map
Heat map
Choropleth map
Proportional symbol map
LoFo An interactive web map application

Fig. 2. Workflow Diagram

Methodology and Application

Data:

- 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.”
- While location-based social media (LBSM) is more than just attaching an instant location or location history to the shared information; it also comprises a new social structure including individuals linked by the exact physical locations or location histories and geo-tagged contents (Zheng, 2011).

Methodology and Application

Data:

- Similarities of VGI and LBSM data: large volume, publicly visible, contain geographic information, and are generated and uploaded by non-specialists.
- 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.

Methodology and Application

Data:

Table 1 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

Methodology and Application

Data:

- 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).
- HLL allows lossless union operation on multiple sets, which makes calculating the number of distinct users (User Count) within each area after aggregation possible.

Methodology and Application

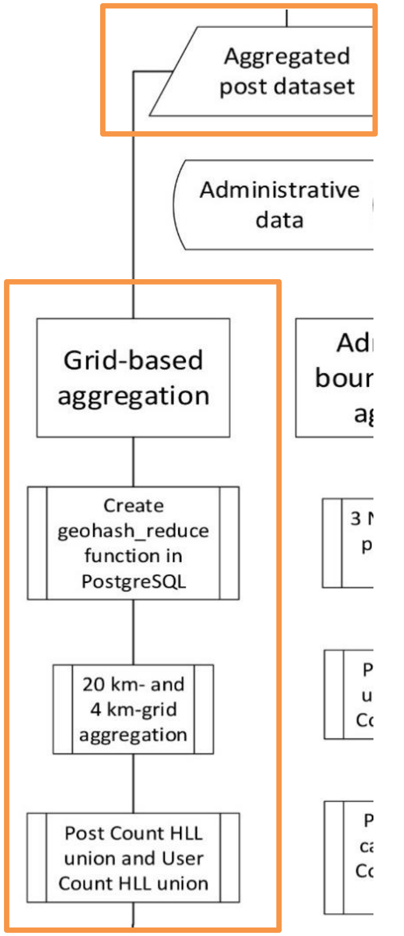
Data Aggregation:

- Grid-based Aggregation
- Administrative Boundaries-based Aggregation
- Density-based Data Clustering
  - Subjective Information Processing

Methodology and Application

Grid-based Aggregation:

- To get an overview of AOIs on small scales which have not high requirements on details, data can be highly aggregated to reduce the influence of the ambiguous geotagged data from social media.
- "width_bucket" function provided by PostgreSQL: allows flexible partitioning of the study area according to needs; it is difficult to apply this method to make the segmented mesh equal in length and width.



Aggregated post dataset
Administrative data
Grid-based aggregation
Create geohash_reduce function in PostgreSQL
20 km- and 4 km-grid aggregation
Post Count HLL union and User Count HLL union
Ad
bour
ag
3 N
p
P
u
Co
P
ca
Co

Methodology and Application

Grid-based Aggregation:

- Inspired by the work of Dunkel et al. (2020), Geohash divides the earth's surface into buckets of grid shape and applying a geohash_reduce function can aggregate spatial data into grids with various sizes corresponding to the Geohash length.

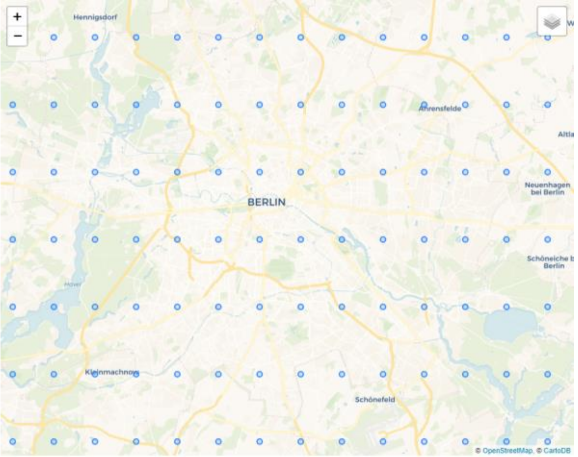


Fig. 3. Data aggregated on 4 km-grids (around Berlin)

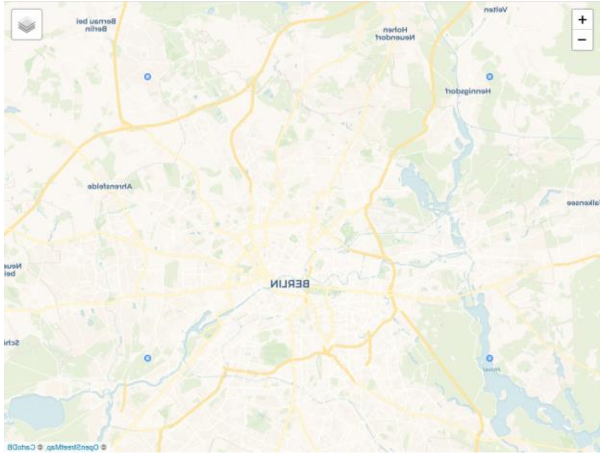


Fig. 4. Data aggregated on 20 km-grids (around Berlin)

Methodology and Application

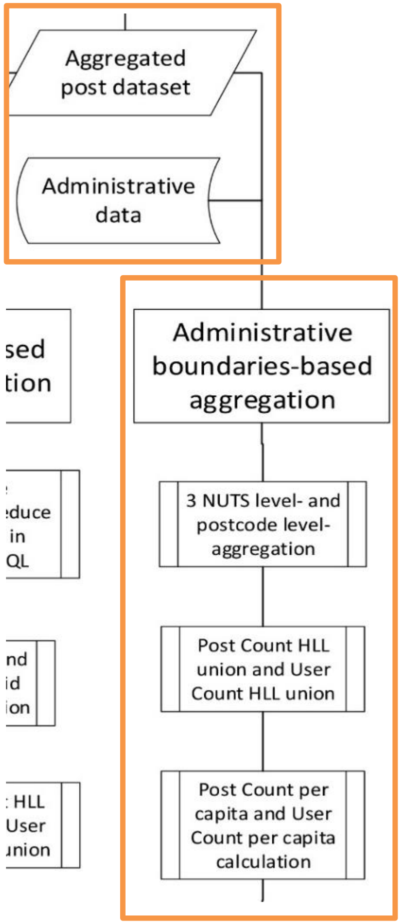
Grid-based Aggregation:

- 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.

Methodology and Application

Administrative Boundaries-based Aggregation:

- 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.
- Nomenclature of Territorial Units for Statistics (NUTS) is 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.
- In Germany:
    - NUTS1: states (Bundesland)
    - NUTS2: government regions (Regierungbezirk, or Direktionsbezirke)
    - NUTS3: Districts (Kreis)


Aggregated post dataset
Administrative data
sed
tion
Administrative boundaries-based aggregation
educe
in
QL
3 NUTS level- and postcode level-aggregation
nd
id
ion
Post Count HLL union and User Count HLL union
HLL
User
union
Post Count per capita and User Count per capita calculation

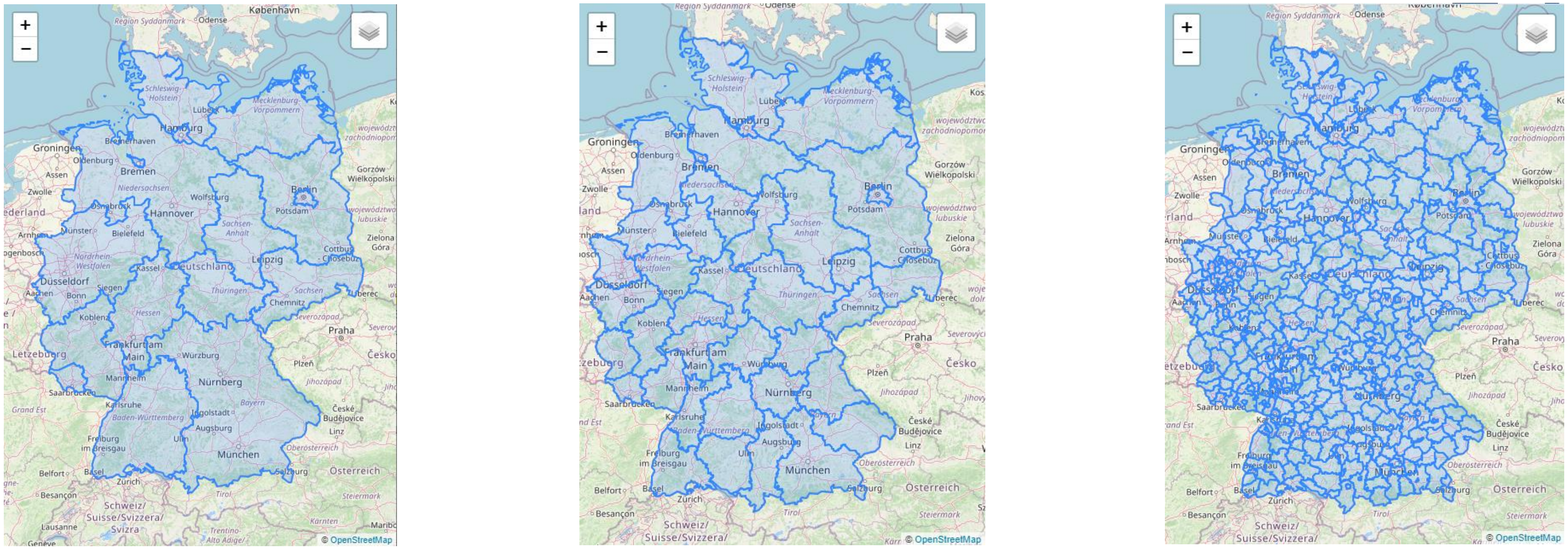
Methodology and Application

Administrative Boundaries-based Aggregation:

- 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.
- Additionally, a country level is added to give a summary of the data in Germany and a postcode level can provides a view within a city.

Methodology and Application

Administrative Boundaries-based Aggregation:



(a) On NUTS1 level

(b) On NUTS2 level

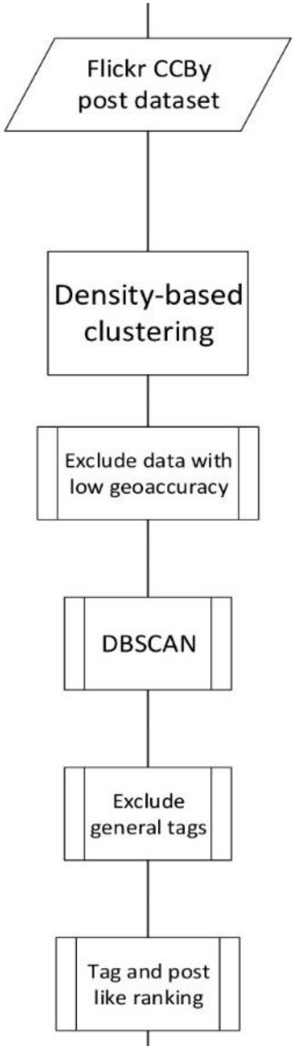
(c) On NUTS3 level

Fig. 5. Data within Germany after NUTS boundaries-based aggregation

Methodology and Application

Density-based Data Clustering:

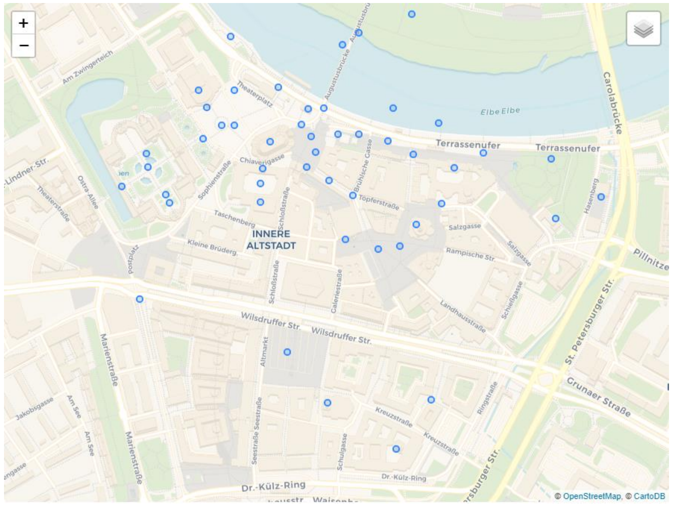
- For extracting POIs on a local level such as famous buildings, parks, or bridges, grid-based aggregation and administrative boundaries-based aggregation are not adequately detailed.
- DBSCAN (density-based spatial clustering of applications with noise): an unsupervised clustering method designed to discover clusters with irregular shapes. Input parameters ε (eps) and minPts can be selected with a good understanding of the data.



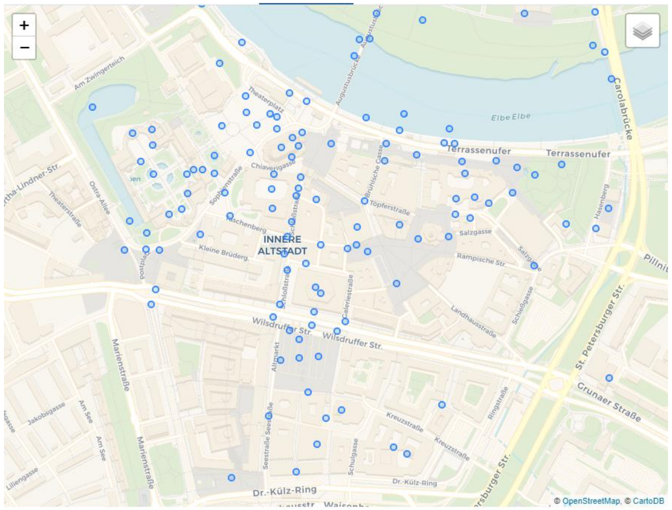
Methodology and Application

Density-based Data Clustering:

- 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.



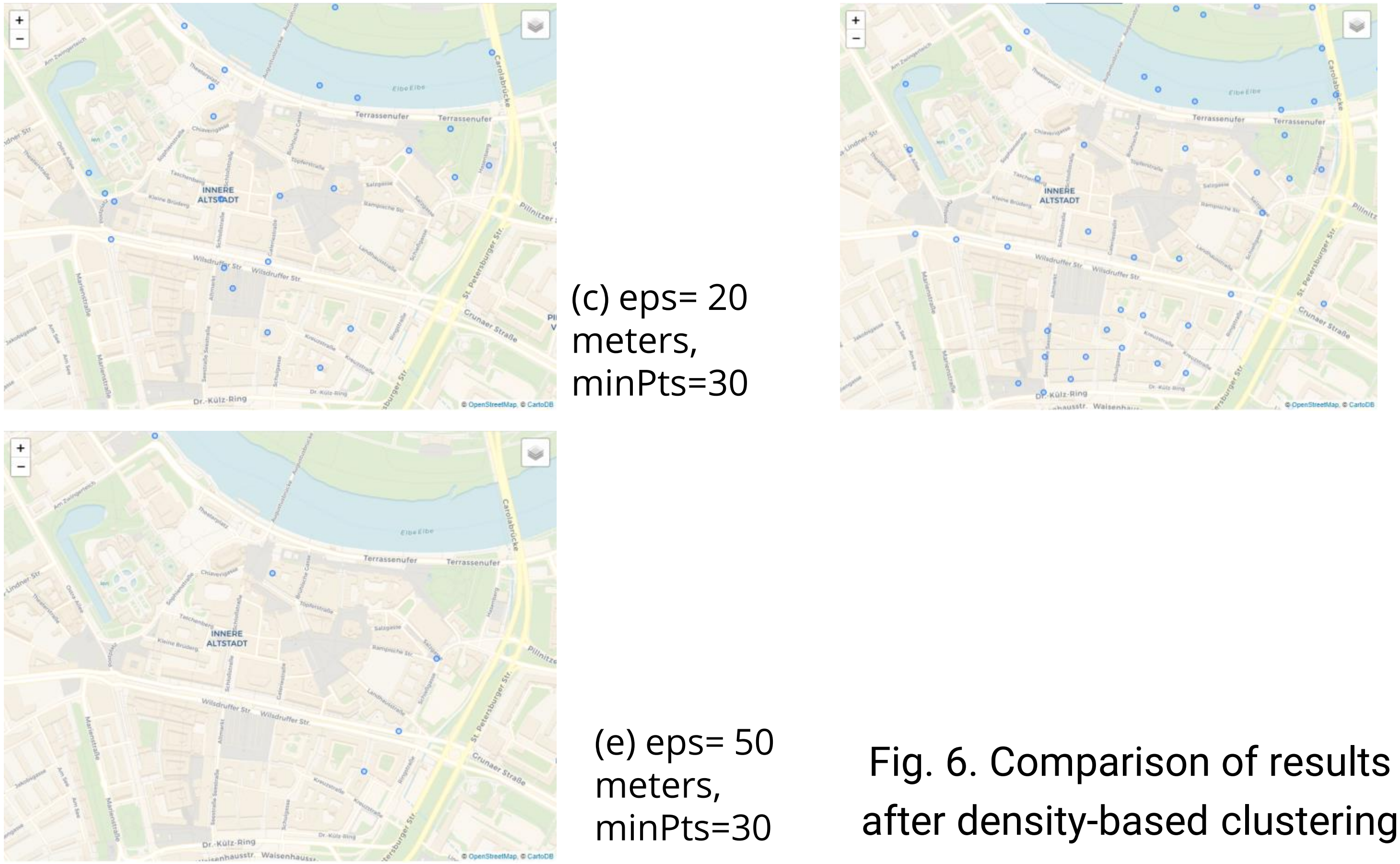
(a) eps= 10 meters, minPts=30



(b) eps= 10 meters, minPts=10

Methodology and Application

Density-based Data Clustering:



(c) eps= 20 meters, minPts=30

(d) eps= 20 meters, minPts=10

(e) eps= 50 meters, minPts=30

Fig. 6. Comparison of results after density-based clustering

Methodology and Application

Subjective Information Processing:

- Besides the popularity information, tourists also tend to be interested in the "content" of a POI. In other words, during the trip-planning or searching for photo-shooting spots stage, the tourists or photographers would like to know what they can see at specific places.
- Tags and thumbnails of photos from Flickr CCBy post dataset are utilized and processed for POI information enhancement and supplement.

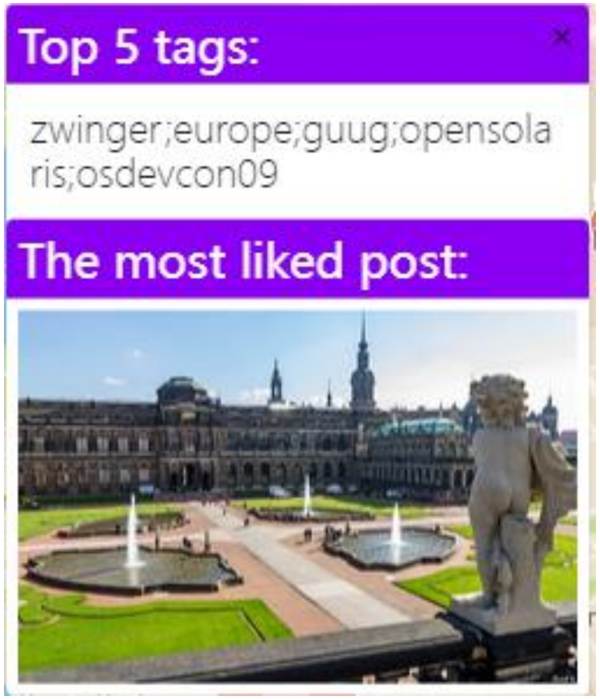
Methodology and Application

Subjective Information Processing:

- 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.

Methodology and Application

Subjective Information Processing:



Top 5 tags:

zwinger;europe;guug;opensolaris;osdevcon09

The most liked post:

Fig. 7. An example showing top 5 tags and the thumbnail of the most-liked photo at a POI

Methodology and Application

Interactive Visualization:

Technologies (tools) adopted:

- Web related technologies: HTML, CSS, JavaScript
- Front-end development framework: Bootstrap
- JavaScript library: Mapbox GL JS
- IDE: Visual Studio Code
- Version management: GitHub
- Deployment: Netlify

Link: https://lofo.netlify.app/

Methodology and Application

Interactive Visualization:

- AOI Heatmap
- AOI Choropleth Map
- Study case in Dresden
- Study case in Berlin

Methodology and Application

Interactive Visualization:

Table 2 Zoom levels

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

Methodology and Application

AOI Heatmap:

- 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.
- While users can switch metrics between postcount and usercount, there are in total four heatmaps varying along with zoom levels in LoFo.

Methodology and Application

AOI Heatmap:

- From zoom level 0 to a maximum zoom level 7, data after aggregation on Geohash length of 4 is applied for the visualization.
- 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.

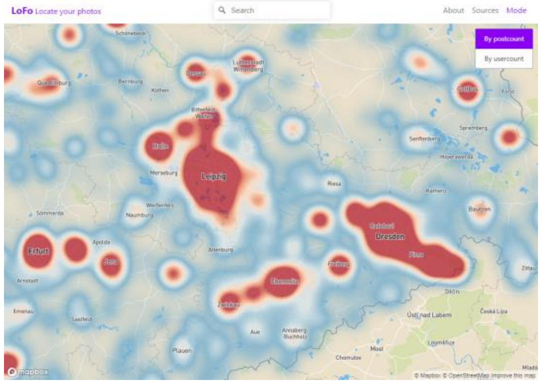


Fig. 8. Heatmap displaying the region around Saxony (at zoom level 8)

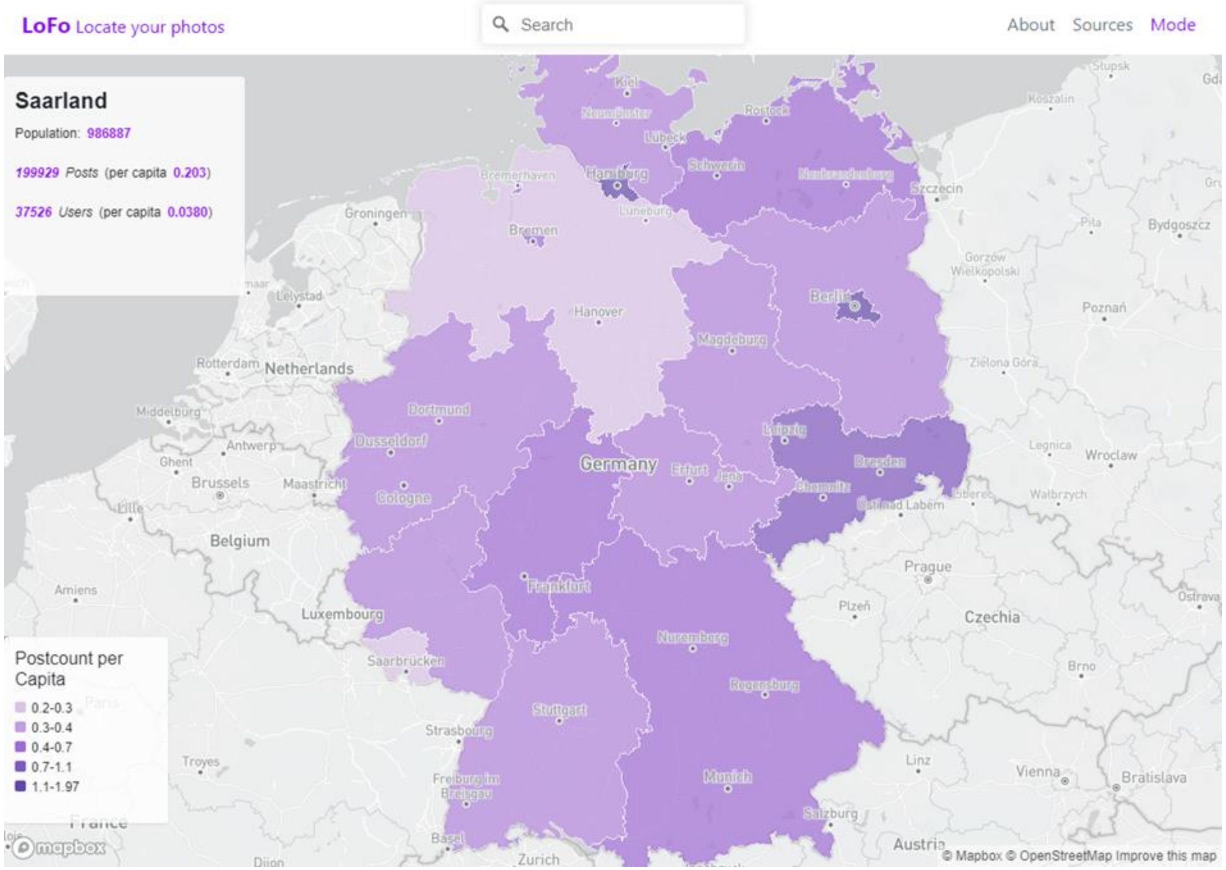
Methodology and Application

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.
  - Zoom level 0-5 Germany
  - Zoom level 5-6 NUTS1
  - Zoom level 6-7 NUTS2
  - Zoom level 7-8 NUTS3
  - Zoom level 8-22   Postcode areas

Methodology and Application

AOI Choropleth Map:



Saarland

Population: 986887

199929 Posts (per capita 0.203)

37526 Users (per capita 0.0380)

Postcount per Capita

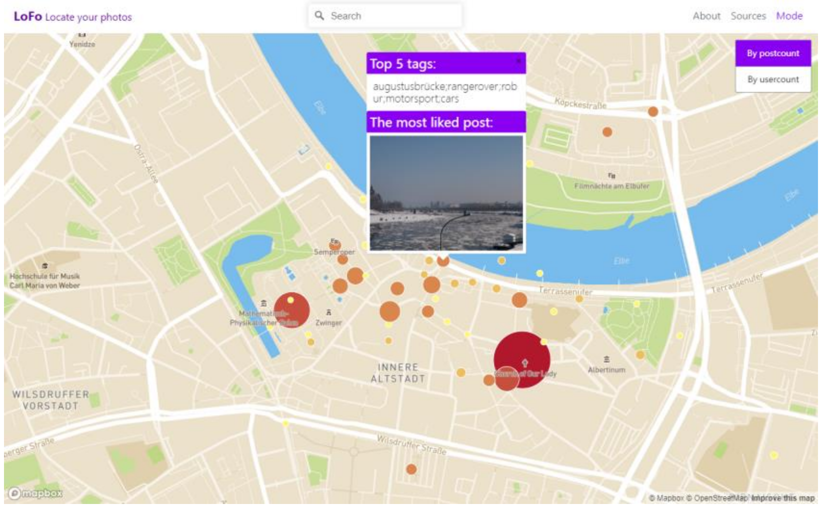
- 0.2-0.3
- 0.3-0.4
- 0.4-0.7
- 0.7-1.1
- 1.1-1.97

Fig. 9. Choropleth map with an aggregation on NUTS1 level

Methodology and Application

Study case in Dresden:

- This map layer in figure 10 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.



Top 5 tags:

augustusbrücke;rangerover;robur;motorsport;cars

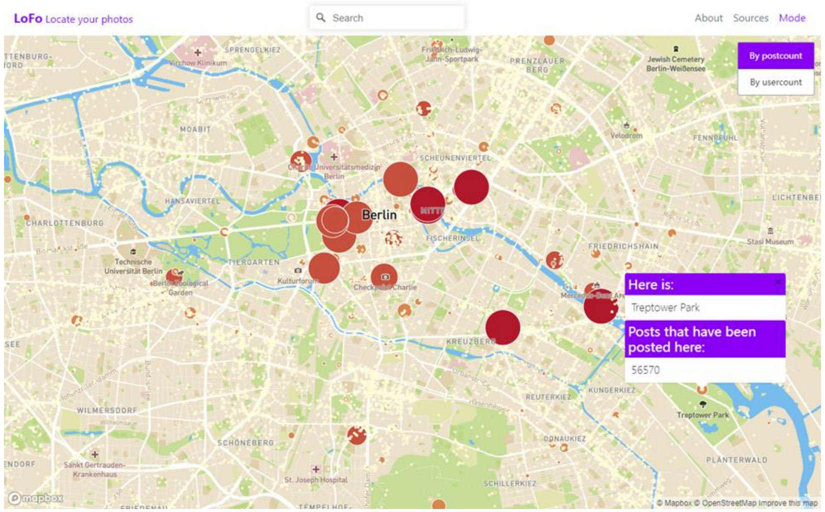
The most liked post:

Fig. 10. Local POIs in Dresden city center

Methodology and Application

Study case in Berlin:

- This map layer, as shown in figure 11, 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 name dataset, and users can check the place name and postcount number of a POI by clicking it.



Here is:
Treptower Park
Posts that have been posted here:
56570

Fig. 11. Local POIs in Berlin

Discussion

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.

Discussion

Evaluation of the privacy protection:

- 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.
- 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.).

Discussion

Evaluation of Data Aggregation and POI extraction:

- Grid-based aggregation using Geohash is highly easy to use, and by creating geohash_reduce function, this type of aggregation is straightforward to implement. Howev-er, 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.
- 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.

Conclusion and Outlook

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 aggerate 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.

Conclusion and Outlook

Outlook:

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

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

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

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

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

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

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

ITC
UNIVERSITY OF TWENTE.
TECHNISCHE UNIVERSITÄT DRESDEN
Technical University of Munich
TUM
TU WIEN
TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology