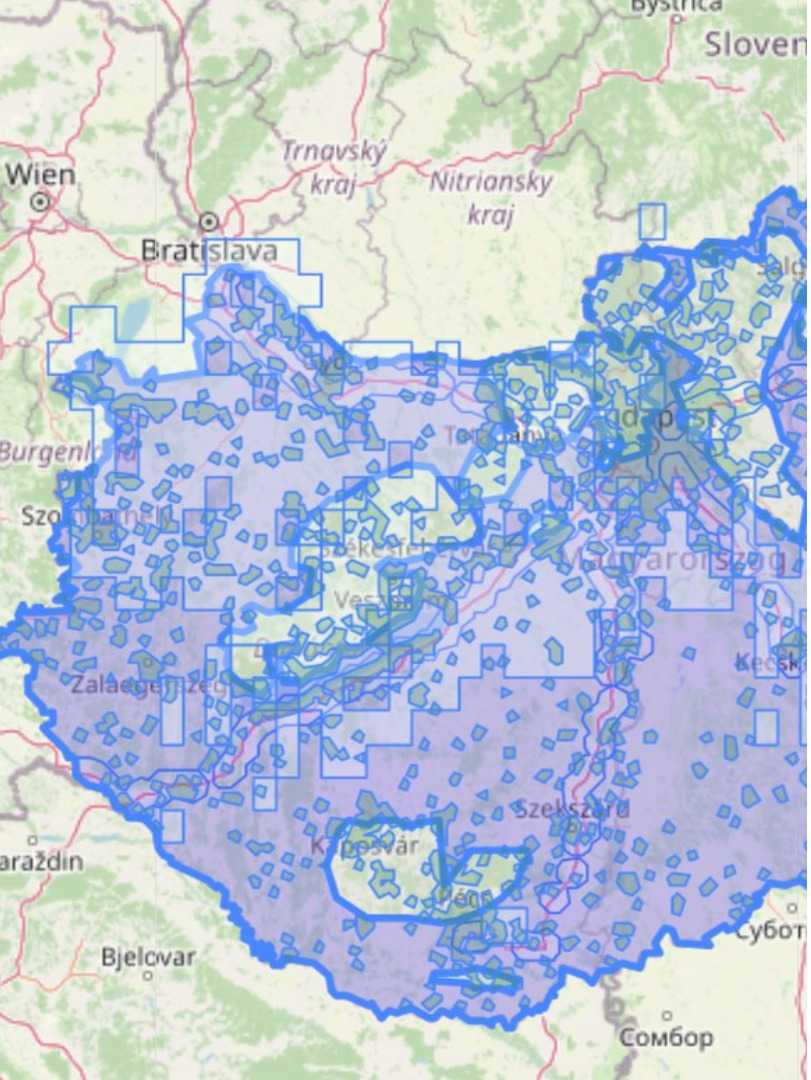
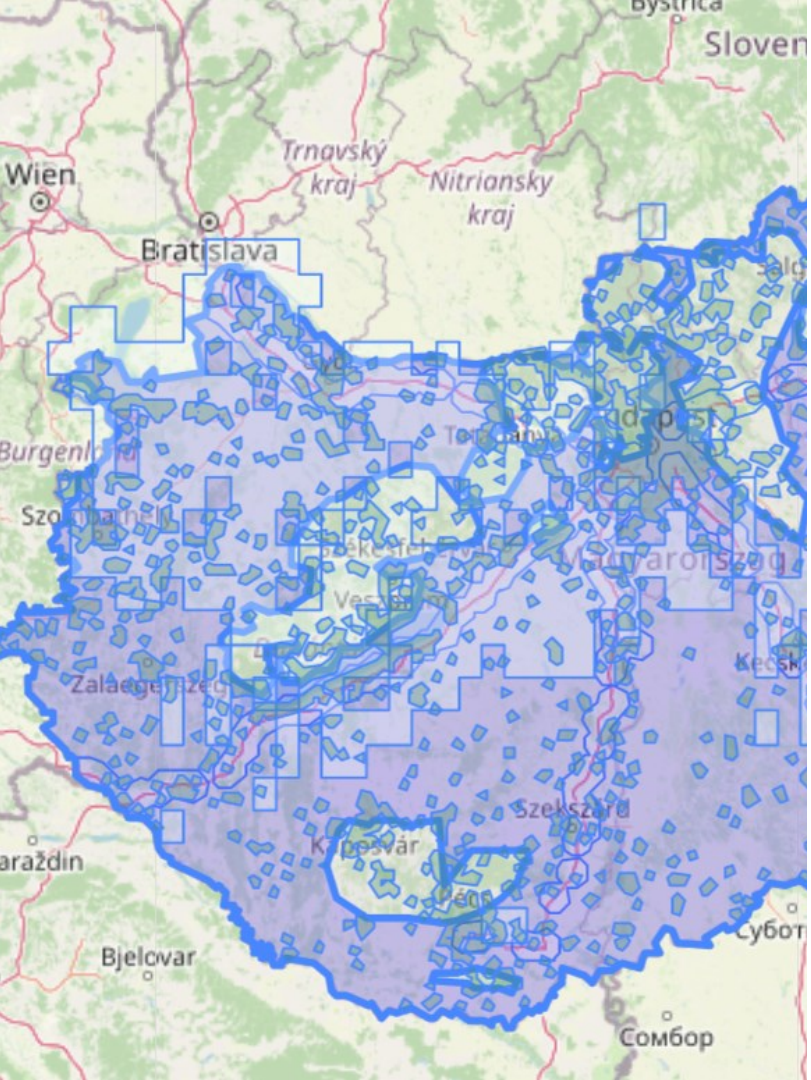



# Manipulation of Geospatial Data: a Dashboard Within Hungary's Geographical Scope

Data Science Bachelor Project
Author: **Gergo Gyori** (gegy@itu.dk)
Supervisor: **Eleni Tzirita Zacharatou**




# Aim: Develop a dashboard that allows users to find specific places based on their constraints

# Geospatial data

## What is geospatial data?

- Data about objects, events, or phenomena that **have a location on the surface of the earth**
    - **Static**: location of amenities, etc
    - **Dynamic**: moving vehicle, the spread of disease, etc
- It combines location information (e.g., coordinates on the earth), attribute information (e.g., the characteristics of the object)
- This project uses **static geospatial data**

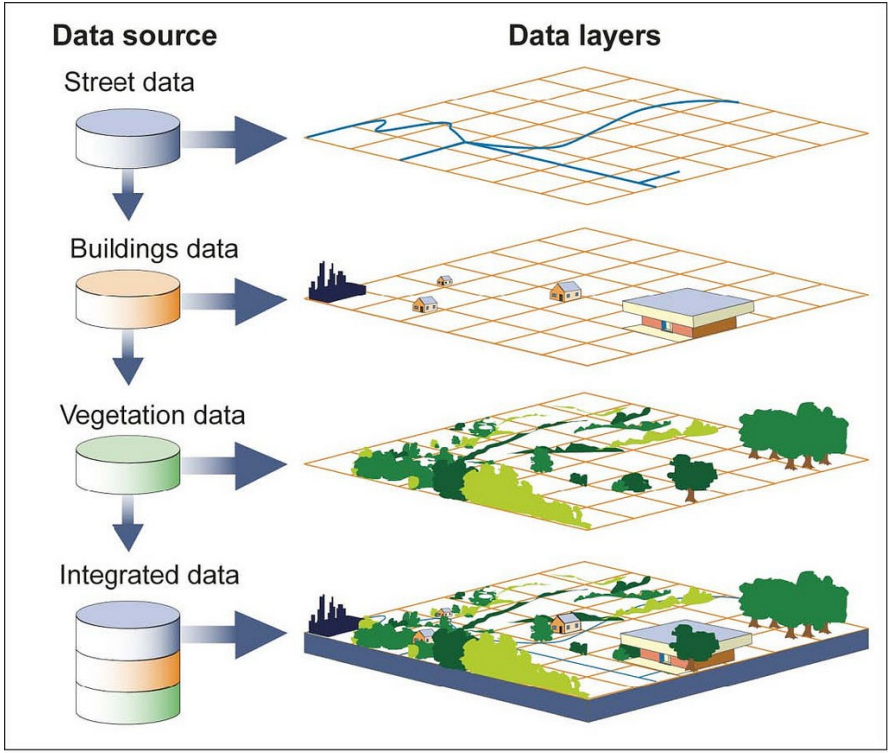



*Image source: https://education.nationalgeographic.org/resource/geographic-information-system-gis/*

# Data source

## OpenStreetMap

- OpenStreetMap (OSM) is a collaborative mapping project that provides a free and editable map of the world.
- The map data in OSM includes a wide range of features such as roads, buildings, parks, rivers, and points of interest
- Open license


OSM: *https://www.openstreetmap.org/*

# OSM - Queries

- OSM key-value tag queries are used to filter and extract specific object from OSM data based on their attribute values

- Each object in OSM is represented by a set of key-value tags

- Users can specify the desired combination to allow **for targeted data extraction**


*A park with green in the middle of the city*


*Example of the "leisure=park" query: results the park objects marked in blue*

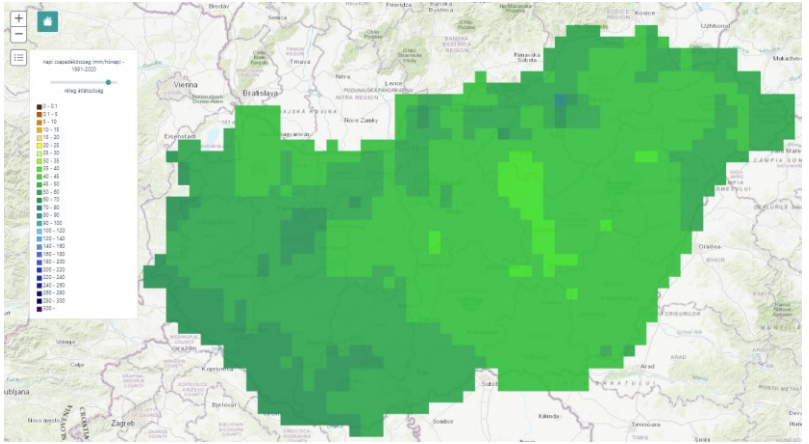
OSM: *https://www.openstreetmap.org/*

# Another data sources

- Klimadat operated by Hungarian Meteorological Service

https://klimadat.met.hu/

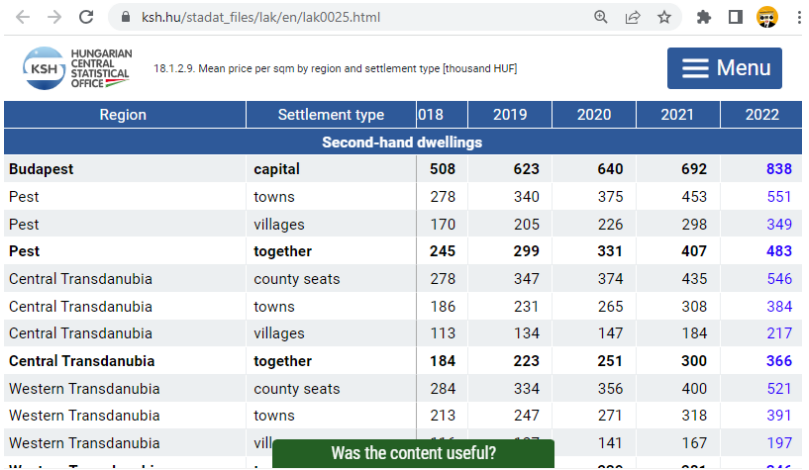
- *Various weather-related data (precipitation, temperature, etc.) and predictions*
- *Based on ALADIN-Climate and REMO models*



- KSH (Hungarian Central Statistical Office)

https://www.ksh.hu/?lang=en

- *Economic, social, and demographic data*
- *Process and analysis of statistical data*



ksh.hu/stadat_files/lak/en/lak0025.html

HUNGARIAN CENTRAL STATISTICAL OFFICE

18.1.2.9. Mean price per sqm by region and settlement type [thousand HUF]

Menu

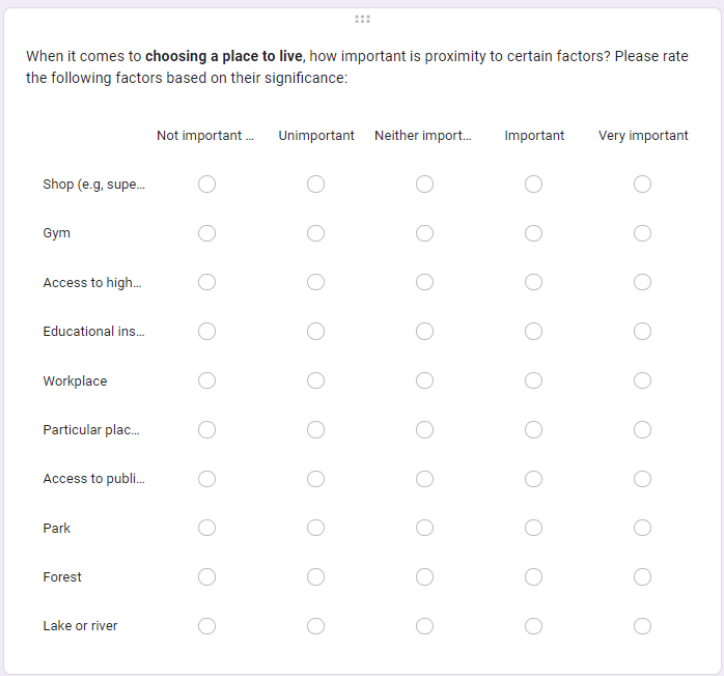
| Region | Settlement type | 018 | 2019 | 2020 | 2021 | 2022 |
|---|---|---|---|---|---|---|
| Second-hand dwellings | | | | | | |
| **Budapest** | **capital** | **508** | **623** | **640** | **692** | **838** |
| Pest | towns | 278 | 340 | 375 | 453 | 551 |
| Pest | villages | 170 | 205 | 226 | 298 | 349 |
| **Pest** | **together** | **245** | **299** | **331** | **407** | **483** |
| Central Transdanubia | county seats | 278 | 347 | 374 | 435 | 546 |
| Central Transdanubia | towns | 186 | 231 | 265 | 308 | 384 |
| Central Transdanubia | villages | 113 | 134 | 147 | 184 | 217 |
| **Central Transdanubia** | **together** | **184** | **223** | **251** | **300** | **366** |
| Western Transdanubia | county seats | 284 | 334 | 356 | 400 | 521 |
| Western Transdanubia | towns | 213 | 247 | 271 | 318 | 391 |
| Western Transdanubia | vill | | | 141 | 167 | 197 |

Was the content useful?

# Users' preferences - Survey

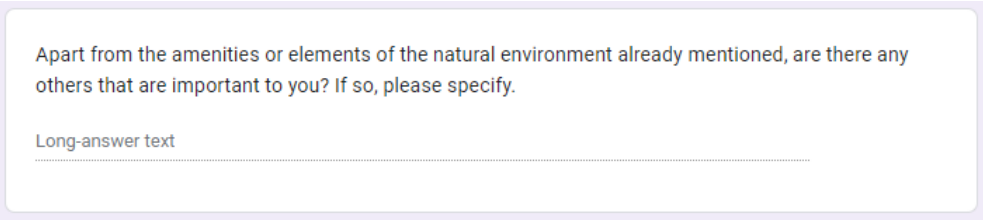


- Free-form and predefined set of question
- 5 points scale
- Focusing on the user's needs
- Divided into two main section:
  - Place to live
  - Place to go on vacation



When it comes to **choosing a place to live**, how important is proximity to certain factors? Please rate the following factors based on their significance:

| | Not important ... | Unimportant | Neither import... | Important | Very important |
|---|---|---|---|---|---|
| Shop (e.g, supe... | ○ | ○ | ○ | ○ | ○ |
| Gym | ○ | ○ | ○ | ○ | ○ |
| Access to high... | ○ | ○ | ○ | ○ | ○ |
| Educational ins... | ○ | ○ | ○ | ○ | ○ |
| Workplace | ○ | ○ | ○ | ○ | ○ |
| Particular plac... | ○ | ○ | ○ | ○ | ○ |
| Access to publi... | ○ | ○ | ○ | ○ | ○ |
| Park | ○ | ○ | ○ | ○ | ○ |
| Forest | ○ | ○ | ○ | ○ | ○ |
| Lake or river | ○ | ○ | ○ | ○ | ○ |



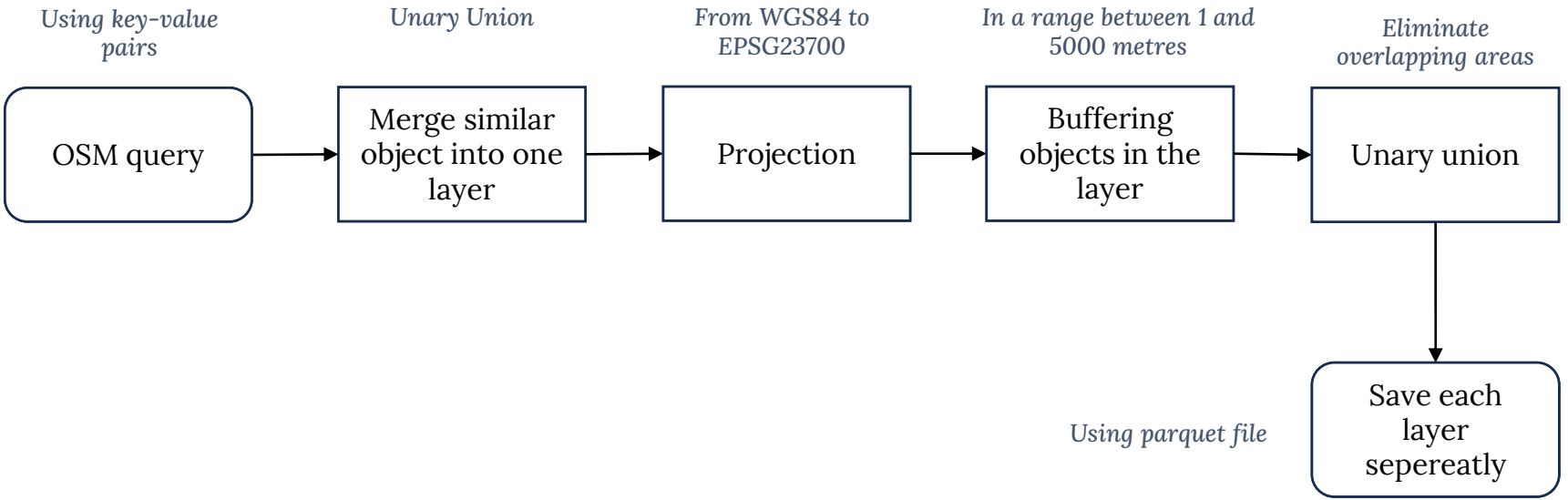
Apart from the amenities or elements of the natural environment already mentioned, are there any others that are important to you? If so, please specify.

Long-answer text

# Flowchart of the pre-processing OSM queries

- Two python packages are used:

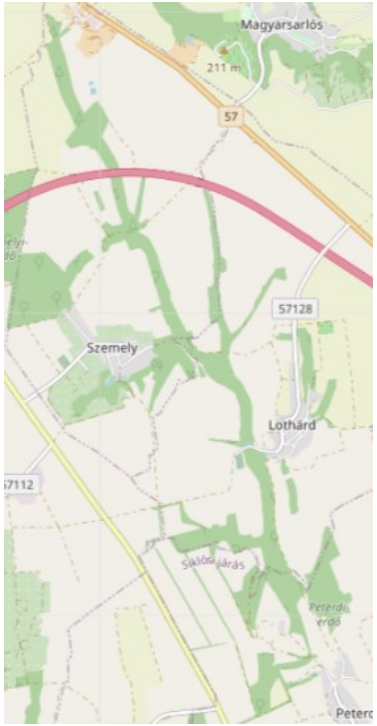
  Geopandas and Shapely



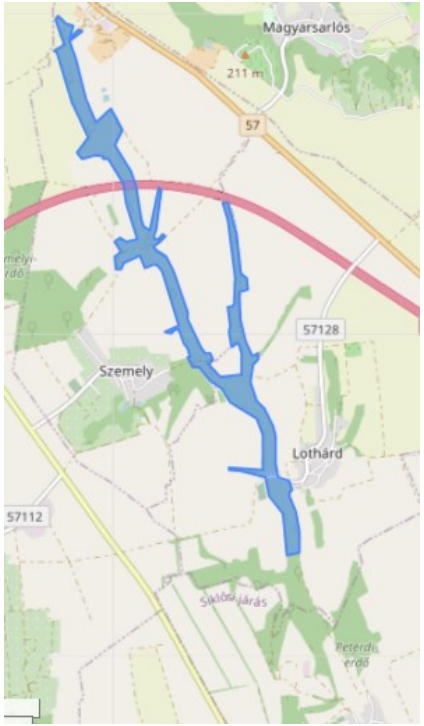
# Unary union

- Improper tagging - merge similar objects
- Eliminate duplicate objects
- Eliminate overlaps after buffering

*A tree branch shape forest area*




*Upper part is tagged as „landuse=forest”*



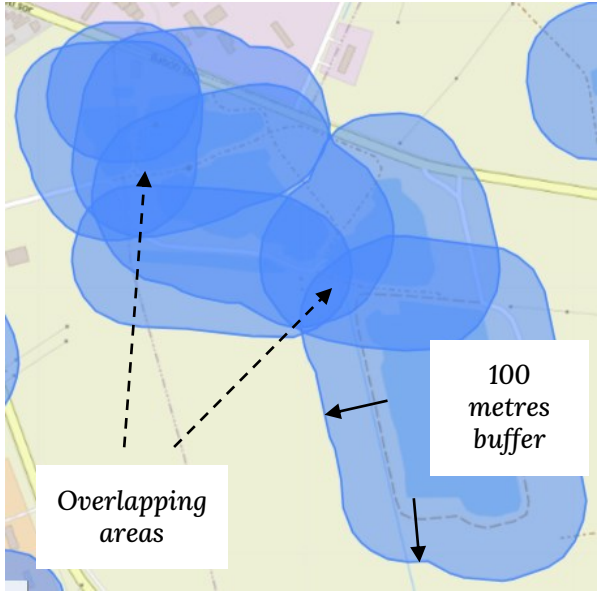

*Bottom part is tagged as „nature=wood”*

# Buffer



*Original objects, some lakes*

*Each objects are buffered by 100 metres*




*Merge into one object after unary union without overlaps*

# Effect of simplification



- No significant improvement in intersection calculation
- Leads to greater inaccuracy
- Douglas-Peucker Algorithm

*Original object*

*Objects is simplified by 10 metres*

| | Result | | | | |
|---|---|---|---|---|---|
| **Simplify value** | **Forest layer area after simp. in square km** | **Water layer area after simp. in square km** | **Number of objects** | **time to calculate intersection (s)** | **Result layer area is square kilometer** |
| 0.1 | 74669 | 32876 | 2711 | 0.47 | 25909 |
| 2 | 74668 | 32873 | 2711 | 0.42 | 25906 |
| 10 | 74594 | 32769 | 2711 | 0.31 | 25801 |

# Result of pre-processing



| Name of the layer | No. of objects | No. of vertices | Total time (buffer, unary union) (s) | Load layer from parquet file (s) | Load layer from CSV (s) | Parquet file size (MB) | CSV file size (MB) |
|---|---|---|---|---|---|---|---|
| Supermarket | 2254 | 53632 | 0.52 | 0.01 | 0.08 | 0.8 | 2 |
| Park | 5573 | 117722 | 1.5 | 0.02 | 0.2 | 1.8 | 4.3 |
| Water | 18999 | 406629 | 11 | 0.03 | 0.65 | 7 | 16 |

*Result table: Objects are buffered by 1000 metres, the layer saved in to parquet or csv file format*

- Time consuming
- It is faster to read the file,

especially in parquet format

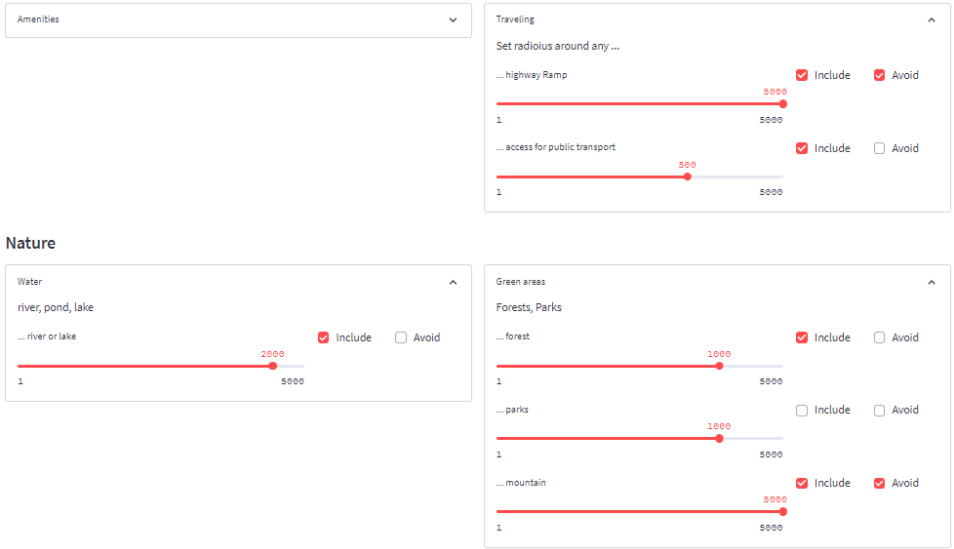
# The dashboard - UI

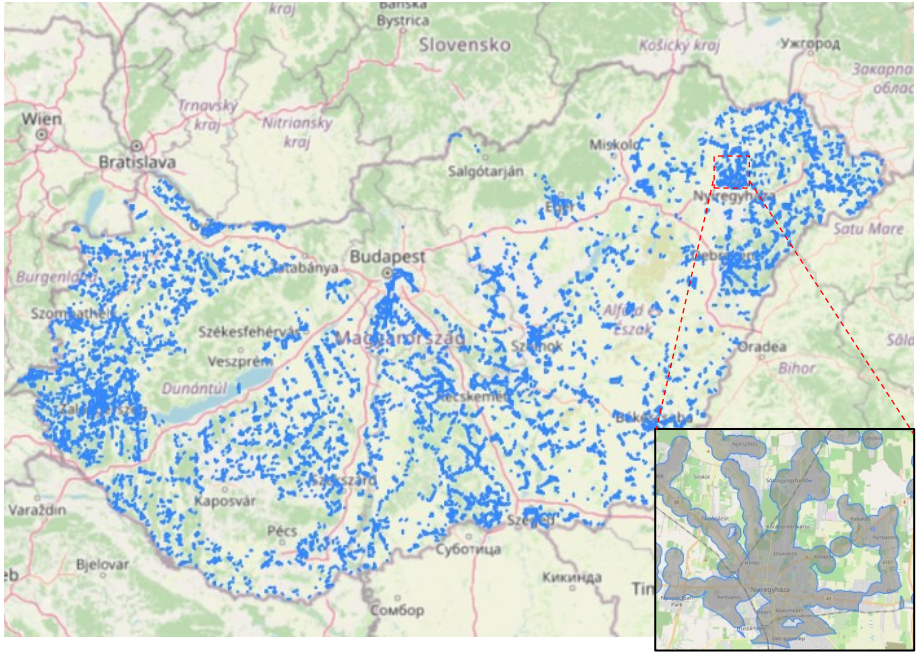


- Hosted on Streamlit

  *https://spot-ideal-places-to-live-in-hungary.streamlit.app/*

*Settings and the result*

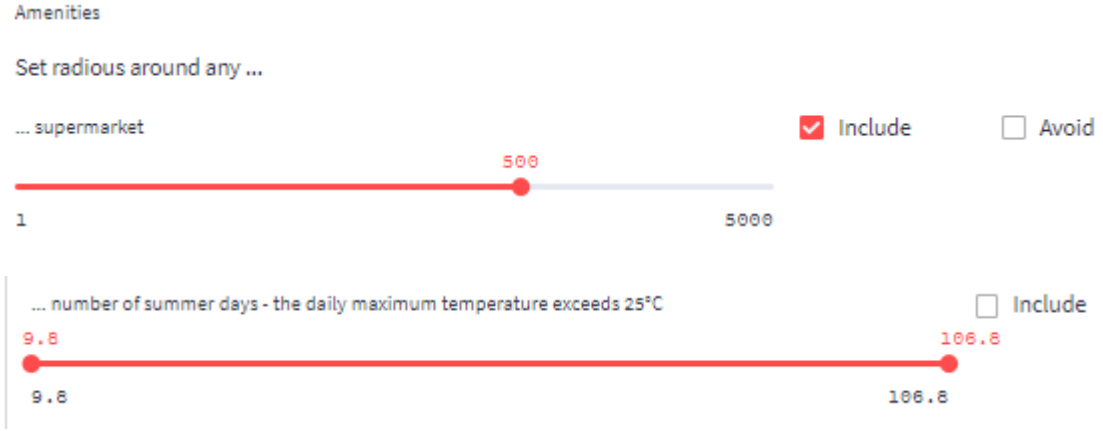



*The structure of the UI and an example result*

# Functions

## Simple sidebar




## Double - ended sidebar

## Avoid


*Areas with blue will be taken into account during the intersection calculation*

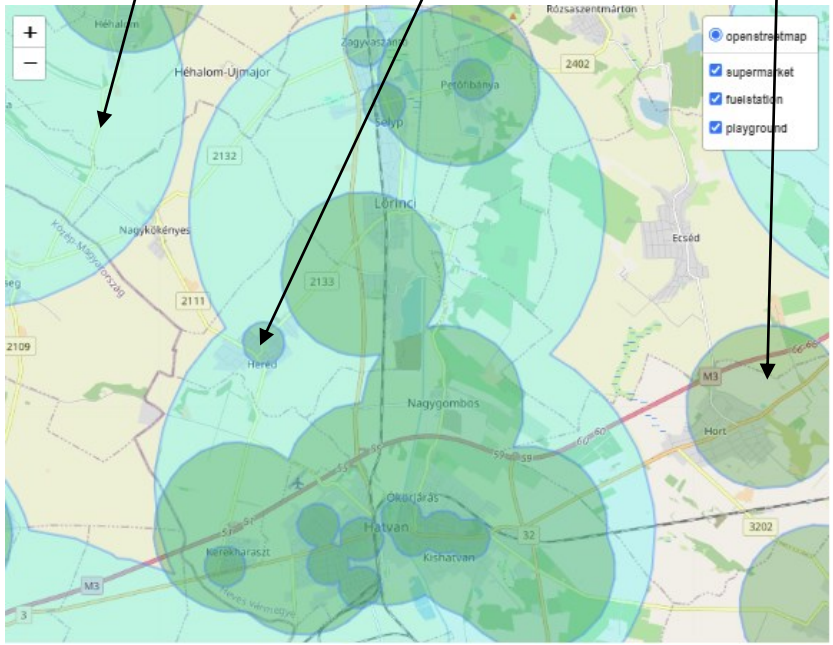
# Functions

## Intersection

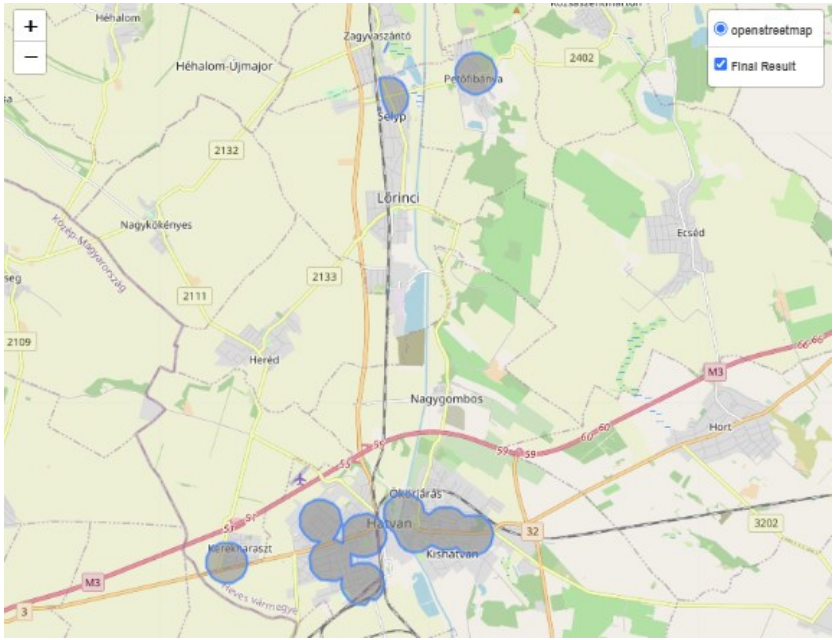
*Blue: supermarket within 5 km*

*Green: playground within 500 metres*

*Green: gas station within 2 km*

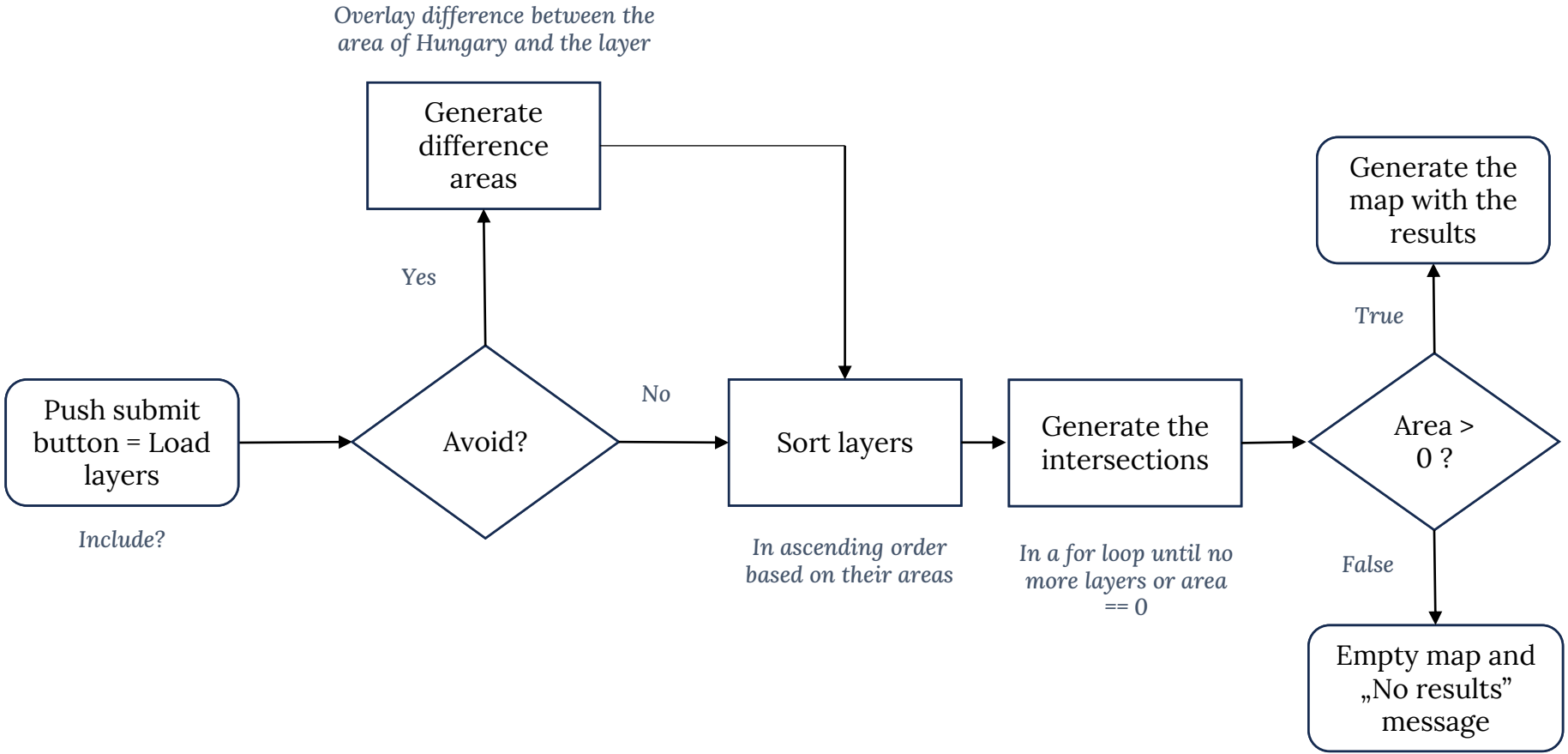



*Result of the intersection of the three layers*



# The back-end of the dashboard





# Limitations, future work, conclusion



- Inaccuracy of OSM database
- Big scale: PostGIS
- The dashboard is available for free for the public
- Can be used by people to find a place to live, agricultural stakeholders, or real estate companies

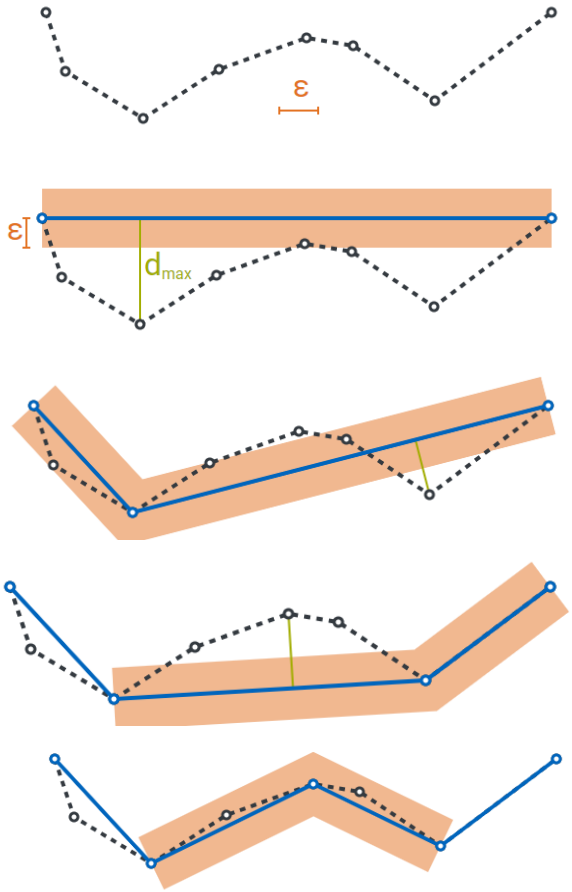
# Thank you

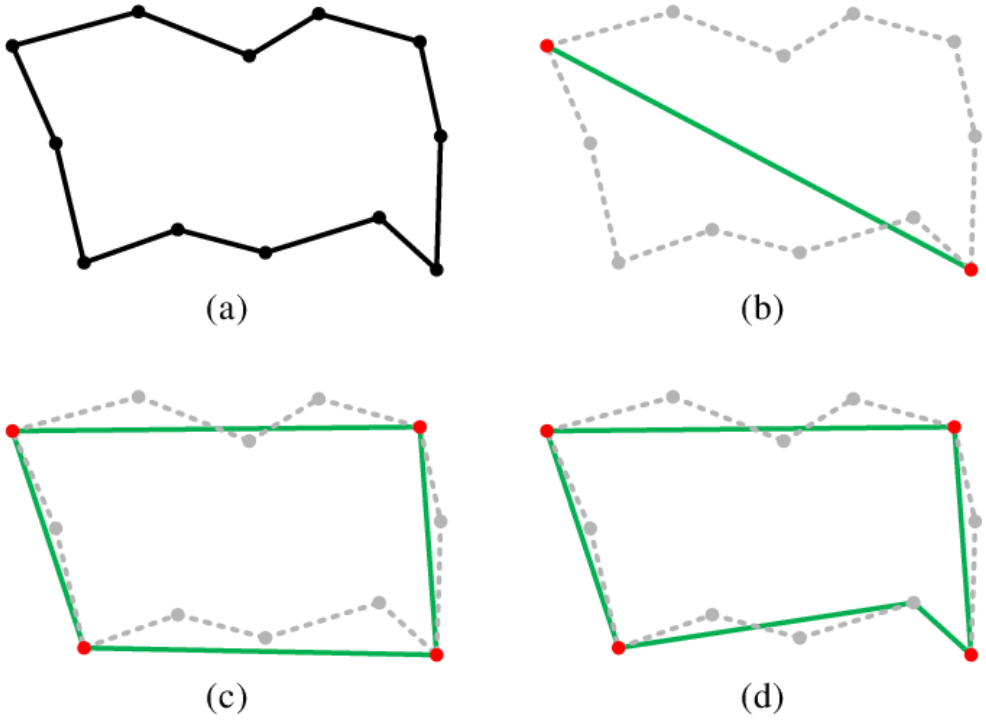
*https://spot-ideal-places-to-live-in-hungary.streamlit.app/*

# Appendix

# Douglas-Peucker algorithm

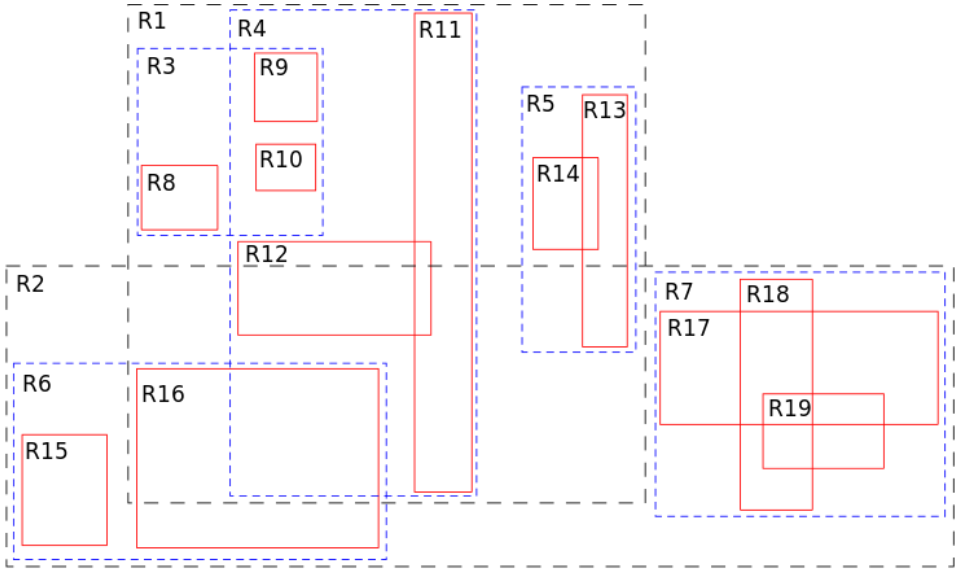


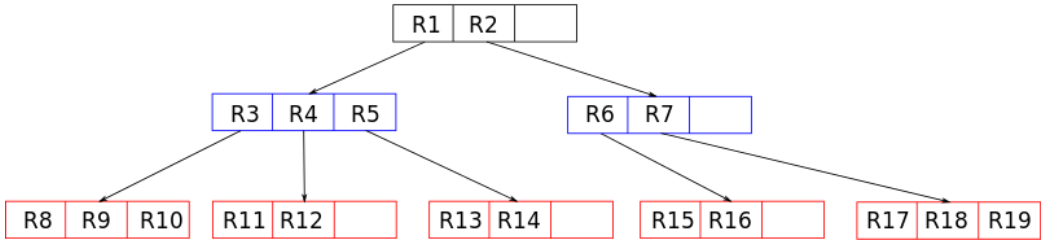


# Indexing – R-Tree

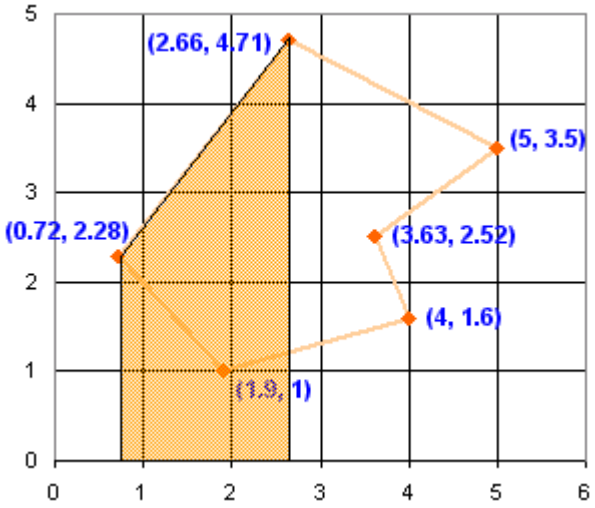




# Area of Polygon

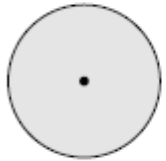




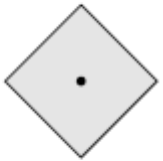

| From | | To | | | | |
|---|---|---|---|---|---|---|
| x | y | x | y | Avg Height | Width (+/-) | Area (+/-) |
| 0.72 | 2.28 | 2.66 | 4.71 | 3.495 | 1.94 | 6.7803 |
| 2.66 | 4.71 | 5 | 3.5 | 4.105 | 2.34 | 9.6057 |
| 5 | 3.5 | 3.63 | 2.52 | 3.01 | -1.37 | -4.1237 |
| 3.63 | 2.52 | 4 | 1.6 | 2.06 | 0.37 | 0.7622 |
| 4 | 1.6 | 1.9 | 1 | 1.3 | -2.1 | -2.7300 |
| 1.9 | 1 | 0.72 | 2.28 | 1.64 | -1.18 | -1.9352 |
| | | | | | Total: | **8.3593** |

# Buffer





Euclidean distance



Manhattan distance

# Clipping