|  |
| --- |
| OTH Regensburg |
| Spatial Databases |
| Project Report |

|  |
| --- |
| Birgit Huber, Florian Fritz, Florian Fußeder  2.1.2017 |

Inhaltsverzeichnis

[Project Summary 2](#_Toc471163020)

[Decisions 2](#_Toc471163021)

[Questions for the User 2](#_Toc471163022)

[Processing the Data 3](#_Toc471163023)

[Neighborhood-Tabulation-Areas 3](#_Toc471163024)

[School Points 3](#_Toc471163025)

[Colleges and Universities 3](#_Toc471163026)

[Parking lot areas 3](#_Toc471163027)

[Population 3](#_Toc471163028)

[Complaint data 3](#_Toc471163029)

[Rental Data 4](#_Toc471163030)

# Project Summary

The goal of the project was to create a web application that helps a user to find a suitable region for accommodation in a city. By answering a series of questions about the preferences and living circumstances we aim to suggest a ranked list of the zones of a city.

Furthermore, the user should be able to enter some main locations. With the given data we are able to calculate a time- and cost-efficient route.

After some research we choose the city “New York” and discussed the which datasets we might use. Using the datasets, we defined our questions for the user and started with processing the data for an easy integration in our project. Afterwards we were able to start with the development of the final webpage.

# Decisions

We started off by considering 9 which we thought might be interesting and might have good datasets. Each member of the team had a closer look at the city and its availed datasets as following.

* Birgit: San Francisco, Beijing, London
* Florian Fr: Berlin, New York, Tokyo
* Florian Fu: Munich, Seoul, Singapore

After a week we had a discussion about the results and choose New York. One of the main reasons for that is that the United States has a lot of open geo datasets because of an open data initiative[[1]](#footnote-1) . New York has its own website with data[[2]](#footnote-2) which we used to get the data for our project. New York is also by default divided into 5 Zones with total area of approx. 790 km² and 194 Neighborhood-Tabulation-Areas (NTA) with an average size of four km² which perfectly suited the needs of our project. With the questions, we wanted to ask the user, we chose to use the following dataset.

* Parks
* Play areas
* Restaurants
* Soccer Fields
* School Points
* Parking lots
* Rental Prices
* Colleges and University’s
* Population
* Complaint Data
* Subways

# Questions for the User

Resulting out of our chosen datasets we picked the following questions to determine a suitable NTA for the user.

* Age range
* Has Children
* Is a student
* Owns a car
* Has a dog
* Does outdoor sports
* Uses subway
* Likes nature
* Prefers vibrant or quiet areas
* Importance of low rental prices
* Prefer to live central
* Favors specific zones

# Processing the Data

Each Dataset was processed individuality. The result of each dataset a database view containing a rating value between zero to one for each region in terms of the data. For example, the area with the least parking lot area scores the lowest rating, the area with the most scores one. This allows us to easily weight the different ratings in the web application as they are now all in the same range.

## Neighborhood-Tabulation-Areas

As mentioned we chose the NTA areas as foundation of our ratings and zones which we want to suggest to our user. The NTA could easily be imported from the exiting shapefiles[[3]](#endnote-1). The Shapefile contained a NTA code, the NTA name, the geometry and some other metadata. Nothing had to be processed in this table.

## School Points

The School point data[[4]](#endnote-2) was imported as shapefile containing locations as geometry based on the official address. It also includes some basic school information such as Name, Address, Principal, and Principal’s contact information.

The rating of how good a NTA is, in terms of School Points, was determined by the amount of points within the area. This was done by using basic postgis queries[[5]](#endnote-3) ,including ST\_AREA and ST\_CONTAINS. As final step a normalized view which only contained NTA code and the rating[[6]](#endnote-4).

## Colleges and Universities

Like the School points table, the data that we used for the colleges also where easy to import because of the shapefile[[7]](#endnote-5) format, and included some metadata like the name and the street name and the geometry.

Again, the rating of areas containing a college or a university is higher than regions that do not. After viewing the data in the web application we came to the conclusion that because of the low number and the bulked locations only a few areas would have a good rating. So we decided it would be better for our rating to use ST\_DWITHIN and add a distance of 200 for our rating view[[8]](#endnote-6).

## Parking lot areas

The parking lot data also already included the geometry of the area[[9]](#endnote-7). So we just had to prepare the data for our rating.

In this case we have two polygon geometries and we want the area of all parking lots so we used ST\_INTERSECTS. To get an accurate rating that is not biased towards bigger NTA areas, we calculated the m² of parking space per m² in the area[[10]](#endnote-8).

## Population

The population data consisted of a simple csv-file[[11]](#endnote-9) that consisted the data from 2000 and the data of the year 2010.

After removing the data from 2000 we had usable data that only had to be match by NTA-code[[12]](#endnote-10). To get a more accurate rating of how crowded the regions are, we calculated the population per m² in the area.

## Complaint data

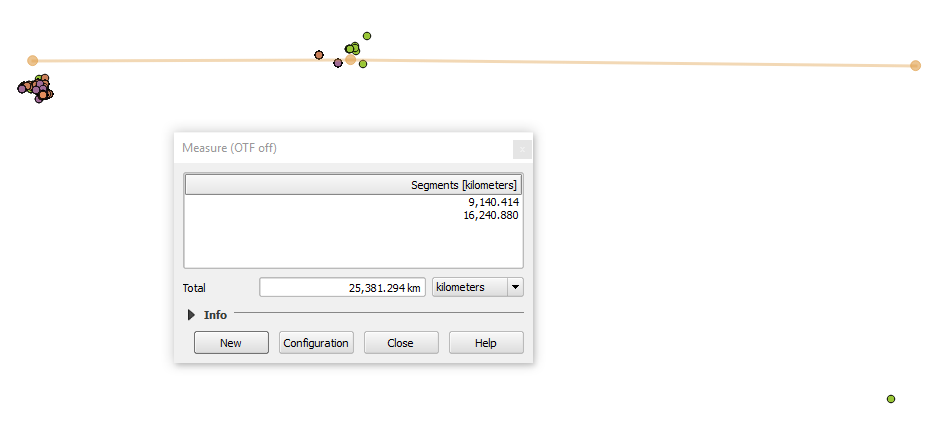
This dataset[[13]](#endnote-11) includes crimes that were reported to the Police for the first three quarters of 2016. The csv-file had the latitude and longitude data some metadata which we cut out to speed up the database queries.

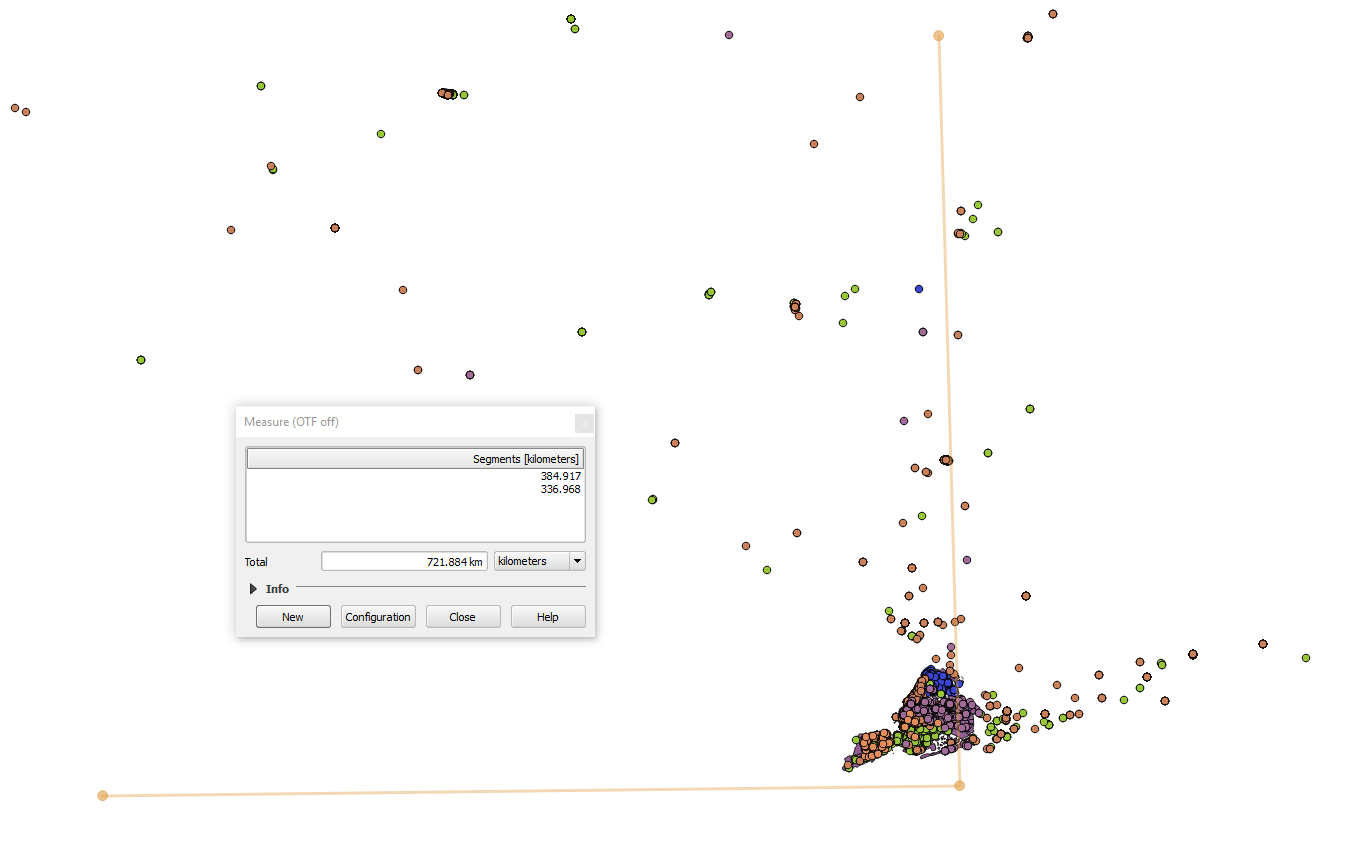
To get a rating of how many crimes have been reported to the police we had to create a point geometry out of the longitude and longitude by using ST\_MAKEPOINT. After that we used ST\_CONTAINS to determine the count of crimes within an area[[14]](#endnote-12).

## Rental Data

To determine a approx. rating of the buying/rental data of the specific areas we used the condominium provided by the NYC department of finances[[15]](#endnote-13).

The datasets consist out of complex Excel Worksheets with much data that where not needed for our project and had to be converted into simple csv-format. Because the data had no spatial data nor a reliable source to match the csv-entries to a specific NTA-area but the address of the building, we decided to use a geocoding[[16]](#footnote-3) API. In our case we used the Python Mapbox API to write a simple script[[17]](#endnote-14) consisting of reading the csv-file, sending the address, receiving the latitude longitude and writing this data into the csv-format. After reviewing the locations of the points that we had produced with the locations from the Mapbox geocoder we discovered that a lot of the produced points where located all over the world. After checking the points, we discovered that approx. 2450 out of the total 23080 entries are clearly out of the boundaries of New York.





Because we decided that we did not want to use such unreliable data, we wrote another Python script[[18]](#endnote-15), but this time we did chose the geocoder API from Google Maps. After checking the locations of the points we created out of the Google latitude and longitude and checking this data in the same way as the Mapbox locations we discovered that only four points where not located in New York. So we did use Google locations to determine average market value per square feet which should give a basic overview of the price range of a NTA.

1. https://www.data.gov/ [↑](#footnote-ref-1)
2. https://nycopendata.socrata.com/ [↑](#footnote-ref-2)
3. <https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas/cpf4-rkhq> [↑](#endnote-ref-1)
4. <https://data.cityofnewyork.us/Education/School-Point-Locations/jfju-ynrr/data> [↑](#endnote-ref-2)
5. <https://github.com/FlorianFusseder/SpatialDatabases/blob/master/FuFlo%20Data/TABLES/4.publicSchoolPoints/SchoolData.sql> [↑](#endnote-ref-3)
6. <https://github.com/FlorianFusseder/SpatialDatabases/blob/master/FuFlo%20Data/TABLES/4.publicSchoolPoints/view.txt> [↑](#endnote-ref-4)
7. <https://data.cityofnewyork.us/Education/Colleges-and-Universities/4kym-4xw5> [↑](#endnote-ref-5)
8. <https://github.com/FlorianFusseder/SpatialDatabases/blob/master/FuFlo%20Data/TABLES/2.colleguesAndUniversitys/view.txt> [↑](#endnote-ref-6)
9. <https://data.cityofnewyork.us/City-Government/Parking-Lot/h7zy-iq3d> [↑](#endnote-ref-7)
10. <https://github.com/FlorianFusseder/SpatialDatabases/blob/master/FuFlo%20Data/TABLES/3.parkintLot/ParkingData.sql> [↑](#endnote-ref-8)
11. <https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp> [↑](#endnote-ref-9)
12. <https://github.com/FlorianFusseder/SpatialDatabases/blob/master/FuFlo%20Data/TABLES/6.population/populationData.sql> [↑](#endnote-ref-10)
13. <https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-YTD/5uac-w243> [↑](#endnote-ref-11)
14. <https://github.com/FlorianFusseder/SpatialDatabases/blob/master/FuFlo%20Data/TABLES/7.complaint/complaintData.sql> [↑](#endnote-ref-12)
15. <http://www1.nyc.gov/site/finance/taxes/property-cooperative-and-condominium-comparables.page> [↑](#endnote-ref-13)
16. Computational process of transforming a postal address to a location (latitude longitude) [↑](#footnote-ref-3)
17. <https://github.com/FlorianFusseder/SpatialDatabases/tree/master/Geocoder/Geocoder_mapbox> [↑](#endnote-ref-14)
18. <https://github.com/FlorianFusseder/SpatialDatabases/tree/master/Geocoder/Geocoder_google> [↑](#endnote-ref-15)