Geospatial Analysis of the Districts of Bulgaria and the City of Plovdiv. Project Description and Results

Milena Atanasova, milena.atanasova@studenti.unitn.it

January 18, 2023

1 Presentation

This project focuses on some exploratory analysis of the country of Bulgaria and of one of its biggest cities Plovdiv, by examining it from a number of different points of view. It provides a descriptive analysis of the districts of Bulgaria (by attempts to see whether some indicator variables should be examined by also considering spatial characteristics), the climate situation in the country, as well as a more concentrated exploration and representation of the city of Plovdiv which is the second largest in Bulgaria after the capital Sofia.

In order to do that, the part concerning the districts is done with the R programming language, while the second is conducted in python notebooks. The project repository can be accessed on GitHub at the following link: https://github.com/atanasova16/GeospatialAnalysisProject.

In order to create the information for the purposes of this analysis, data has been collected from several sources, all of it is described in the sections below.

2 Districts analysis and possible spatial spillovers

2.1 Data

The data for this part of the project comes from two sources. One is the Bulgarian National Statistical Institute [1]. A dataset was built collecting data from there which is freely available. Datasets where information was divided according to NUTS 3 specification was taken for inspection. The variables are: water used by households from public water supply; public water supply; generated municipal waste; accommodation establishments and health establishments[1].

Three of the variables present in the dataset instead, were computed and added from the Python part of the project, by using climate data from WorldClim [2]. From there, average temperature, precipitation and wind speed were extracted, then with some raster exploitation, the tiles were divided according to the Bulgarian NUTS 3 classification, and average was taken for each region, then put in a dataframe, which is then merged with the other information, together with a shapefile representing the geometries of the above mentioned regions (as polygons).

2.2 Analysis

In order to see if different indicator variables can be explained partially by their geographic location, some spatial autocorrelation analysis is conducted, both concerning global and local possible effects. Then, models were tried in order to check the dependence of GDP on some of the variables mentioned above, and also the dependence of average temperature on wind speed and precipitation. These were tested for the existence of spatial dependence as well by using the Lagrange multiplier (LM) test of spatial dependence on OLS residuals.

Firstly, weight matrices are defined according to different specifications. They serve as spatial weights for defining some kind of *spatial similarity*. Then Moran's I test was tried in order to check the existence of such a dependence, however p-values were quite high, so it does not seemed to be significant. The test was also applied to a specification of a linear model with the climate variables.

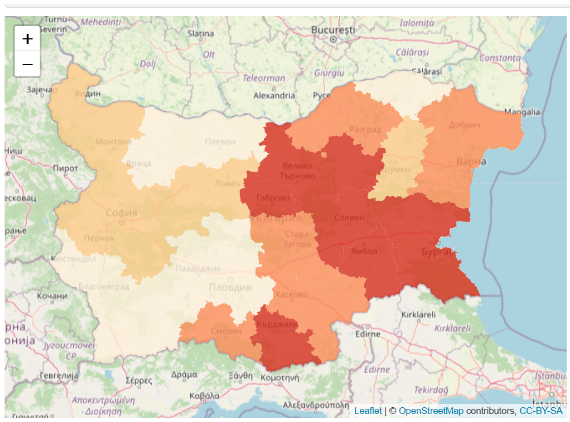
[1] The data refer to the year 2020 since this is the latest information available for the variables

The model estimates (even though the number of districts is 28, therefore the sample size is not very big, the results still can tell something) show that wind speed and rain have a negative effect on the temperature in the districts.

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           32.13380    3.31975   9.680 6.16e-10 ***
precipitation_average -0.32500    0.05355  -6.069 2.43e-06 ***
wind_speed_average    -2.16096    0.51003  -4.237 0.000269 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8301 on 25 degrees of freedom
Multiple R-squared:  0.6327, Adjusted R-squared:  0.6033
F-statistic: 21.53 on 2 and 25 DF,  p-value: 3.654e-06
```

The figure below plots the OLS residuals. There could be seen some kinds of similarities, but not too big. And after statistically checking, it was concluded that they not really significant even though the Moran's statistic is exhibiting a bit lower p-values. Therefore, a global effect could not be found at this stage.



Following, local tests and plots were also conducted, but again, nothing showing any significant results of presence of spillover effects.

Simply to conclude this analysis, it is checked if there exists the need to include spatial characteristics when estimating models. For the purpose, the Lagrange multiplier (LM) test of spatial dependence on OLS residuals is utilized, which offers a way to compare models with and without spatial specifications. After having tried with different weight matrices, it seems there is no presence of importance of the spatial distribution for the models about average temperature and GDP.

From the analysis conducted, it appears that overall in Bulgarian districts, the geographical location does not play a big role for some statistical indicators, as well as the climate variables. The local/global effects cannot be seen, and in fact since the main and most developed cities of the country are more or less scattered around the country, it makes sense.

Therefore, if regression is to be done on some factor concerning the variables at hand, it can be relied on linear regression models with OLS estimator.

There could be some other indicator variables for the country, which could possibly possess the need to be analysed with the use of a spatial regression model.

In fact, if analysis is done based on municipalities, some more useful and informative results could be obtained, also due to the presence of a larger number of statistical units. However, not much data divided according to these NUTS 4 specification is available for the ccountry.

3 The climate in Bulgaria

3.1 Data

As mentioned earlier, climate data comes from WorldClim (in a *GeoTiff* format, at a 2.5 mins spatial resolution for average temperature (in degrees celcius), and at 5 mins for wind speed (in m/s-1) and precipitation (in mm)). The data is historical and is for the whole planet, divided by months.

3.2 Analysis

The data is in the form of a raster layer, therefore some aggregation had to be done in order to get numbers which are representative if it is to be grouped by districts, which seems reasonable to do. In such a way, they can be compared and combined with other data available.

In order to examine the particularity of the climate in Bulgaria, three main activities were conducted, namely:

- **Extracting average temperature/rain/wind from the twelve months for each district**

  The procedure followed was to go through the rasters of each month with all the information for the whole world, then masking with the geometries of each district, and computing the average value from the cells contained in this clipped variant, representing the district. These values, together with the districts' names and geometries, were put together in a single geodataframe that is available in the project repository. Those were the values used for the districts analysis. Below, the three choropleth maps represent this information in a visible way.

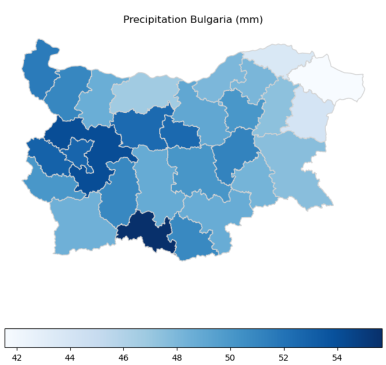


Figure 1: Precipitation

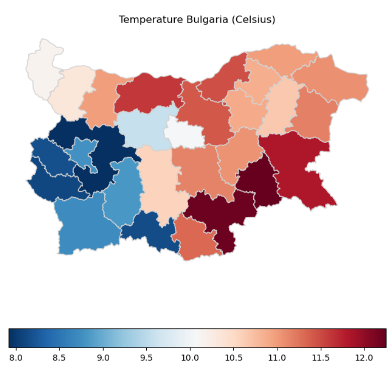


Figure 2: Temperature

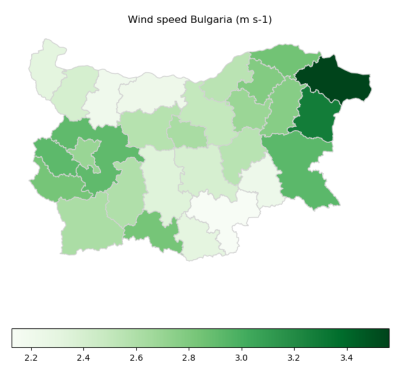


Figure 3: Wind

- **Finding the coordinates of the points which have the highest values for average temperature/rain/wind**

  After having how the tiles' values combine to create an average for the city, it is then proceeded with identification of the places with the maximal value for these averages, and it is done for each district.

  This could be useful if it is to be analyzed for example where some renewable energy plants could be constructed (if wind is considered, decide on places to put wind turbines and estimate how much energy it would be possible to collect). This kind of information may be also utilized for rain water collecting, or to decide on where and how to plant specific kinds of vegetation, also crops.

  Since the spatial resolution is not exactly points, the values are not exactly just points, but represent small regions, where, as proposed, could be thought of building facilities.

  From the map, it is evident that places which have a lot of rain and wind are usually in the mountains (exceptions the north-eastern parts, and on the Black Sea cost, which are actually places known for their wind). Instead, hottest averages tend to be deep in the valleys and in the plains, near cities.

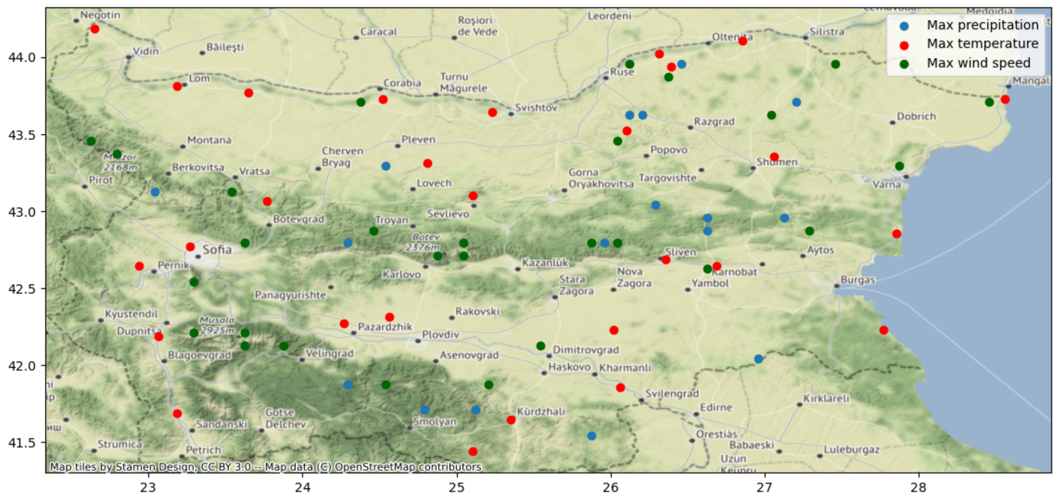


Figure 4: Map with coordinates of points with maximum values for the three climate variables considered

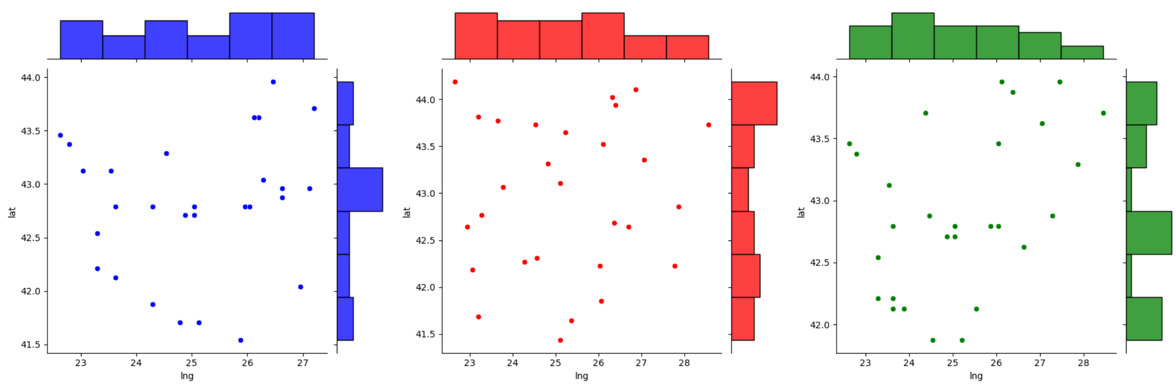
- **Examining the spatial distributions of these points**in order to get a better perspective of where they are located, the heatmaps and the joint plot can provide some visual evidence.

  An essential thing to note is that since by design there are points for each district, they are scattered all around the country.

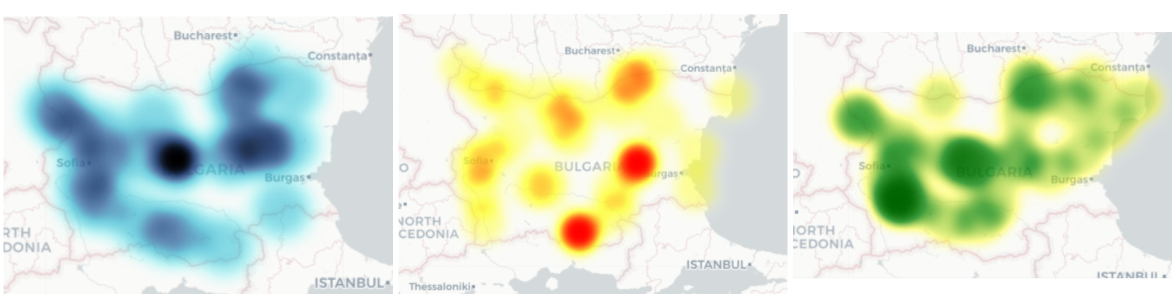
  Even though the points are not too many, a general idea is presented.

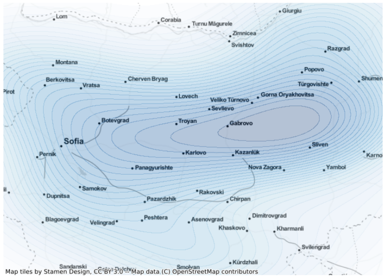
  Blues refer to precipitation, reds to temperature, and greens to wind.

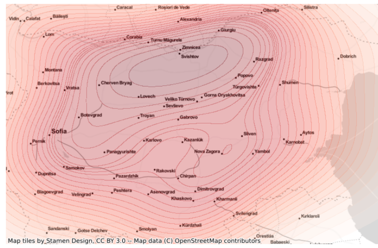
Firstly, the jointplots can tell us for example that rain in the mountain range cutting the country in two is of particularly high value. Wind has three main locations: the same mountain range, near the Black sea coast, and in the south-western high mountains.

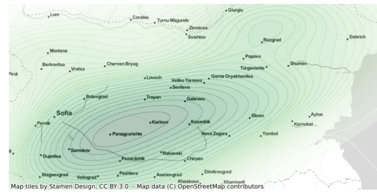


It can be better seen from the heatmaps, two types of which are exhibited below. They show a more fluid representation of the points, which is useful, also in order to see how regions connect to each other for such kinds of maximum values.









Since the information is at district level, each governmental body could possibly consider its own sources, but still a global initiative could be feasible as well. This weather information could be combined with those from the following section, coming from the satellites, and in particular with the NDVI, Agriculture, and Moisture parts. This is a possible point of further development of this project.

3.3 Climate of Plovdiv

Having examined those issues, it is then turned to look in more details about the city of Plovdiv. In fact, as a connection part, below can be seen two maps created with QGIS which show the aforementioned rasters with average values for the districts for Plovdiv.

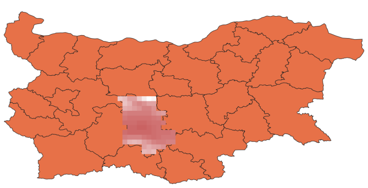


Figure 5: Precipitation

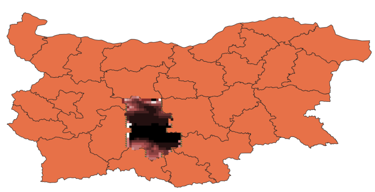


Figure 6: Temperature

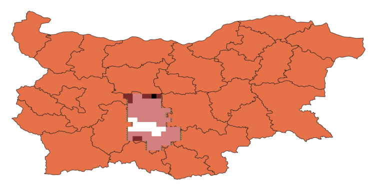


Figure 7: Wind

During the analysis, a certain aspect was to see how the climate is in Plovdiv, and the next three diagrams show the averages for Plovdiv city and Plovdiv district for each month. Overall, in the town, where there is less percentage of nature overall, there is less rain and wind, and is usually warmer. IN fact also the infrastructure plays a large role for these variables.

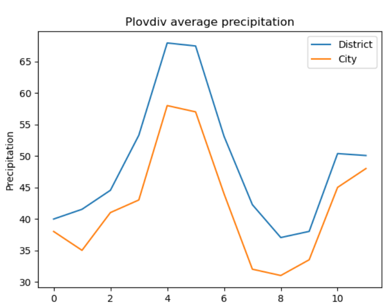


Figure 8: Precipitation

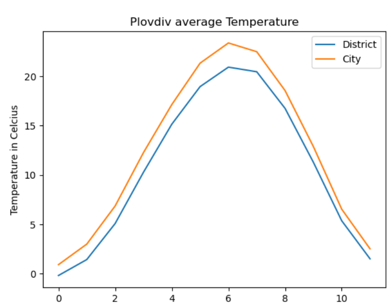


Figure 9: Temperature

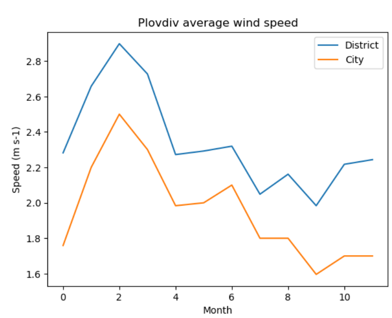


Figure 10: Wind

Still, for the next sections, some more information is taken for examination.

4 An Exploratory Analysis of the City of Plovdiv

The analysis is continued by turning the focus to Plovdiv, as well as using other type of data as well.

4.1 Data

There are three essential data sources for this part of the analysis.

One is the Copernicus Open Access Hub [3] [2], which offers numerous data resources, and for this project, materials from ESA's Sentinel-2 satellite were taken for work. The date come as rasters, again divided in tiles, and one can filter the results requested on amount of cloud coverage available. For this project, data with minimal cloud is being used. The satellite carries the Multispectral Imager (MSI). This sensor delivers 13 spectral bands main bands, and these layers range from infrared to ultraviolet (at 10, 20, and 60 metres spatial resolutions)[4]. The bands obtain light with different wave length, and from that much information could be extracted (especially if combined with more domain knowledge in natural sciences).

Secondly, Digital terrain model was also extracted from the NASA's Earthdata repository [5] [3] which comes from the Shuttle Radar Topography Mission (SRTM).

Finally, a protocol buffer file with to use for OSM representations was exported using BBBike data[6].

All of the above mentioned data resources were extracted with regards to the coordinates of Plovdiv. For the rasters, the tiles which contain a bounding box around the coordinates of Plovdiv were used (and then for the analysis, are clipped), and for the protocol buffer, by defining a bounding box of search.

4.2 Analysis

4.2.1 Satellite output analysis and indices computation

Next are presented the different indices and combinations of different layers of the satellite outputs, which serve different purposes [7].

The Normalized Difference Vegetation Index (NDVI) is a numerical indicator that uses the red and near-infrared spectral bands. NDVI is highly associated with vegetation content. High NDVI values correspond to areas that reflect more in the near-infrared spectrum. Higher reflectance in the near-infrared correspond to denser and healthier vegetation.

The moisture index is ideal for finding water stress in plants. It uses the short-wave and near-infrared to generate an index of moisture content. In general, wetter vegetation has higher values. But lower moisture index values suggest plants are under stress from insufficient moisture.

In order to compute the moisture, and also to represent the next figures, some rescaling had to be done due to the difference in the spatial resolution as stated before.

[2] ESA's Copernicus mission (Contains modified Copernicus Sentinel data [2022])

[3] Shuttle Radar Topography Mission (SRTM) - U.S. Geological Survey

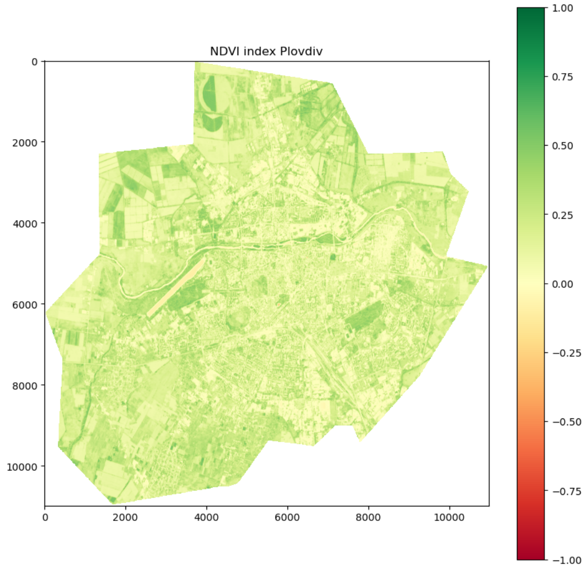


Figure 11: NDVI

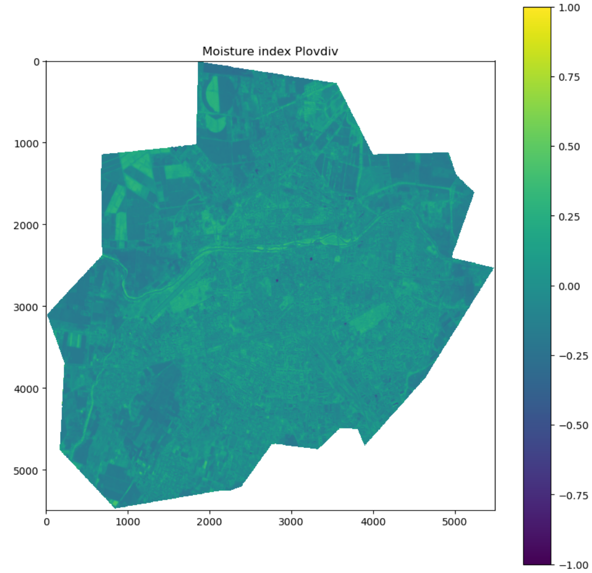


Figure 12: Moisture

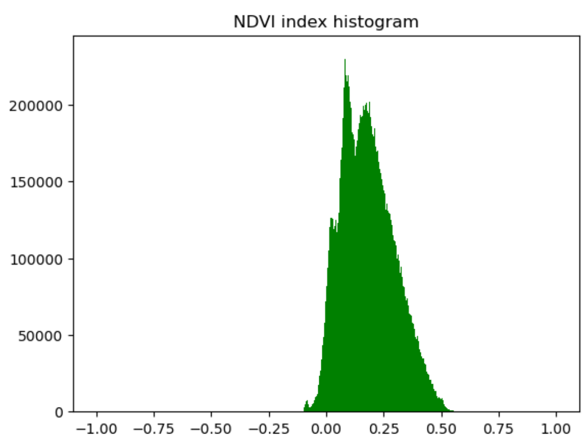


Figure 13: NDVI histogram

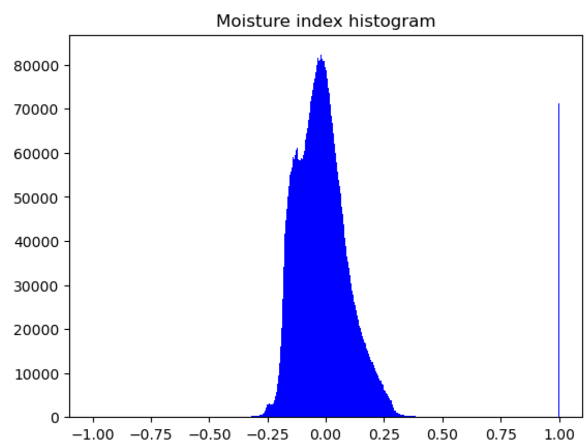


Figure 14: Moisture histogram

The two histograms show the distribution of how the two indices are represented in the city of Plovdiv, and it can be noted that overall they look like normal ones, even if not exactly (two small peaks for NDVI). Another thing to look at is the scale. Overall NDVI is not particularly high, but still in farmlands it appears vegetation index and moisture are not very bad.

Such things are important for a big city, as are crucial for a healthy life. In fact, Plovdiv is a place which has a very high degree of air pollution, so finding ways to have a healthy vegetation (even if not solve the problem in any case) would be good for the environment and those living there.

The next figures show some other kinds of representations.

Agriculture should be able to look at crops.

As the name implies, the bathymetric band combination is good for coastal studies. The bathymetric band combination uses the red (B4), green (B3), and coastal band (B1). Using the coastal aerosol band is good for estimating suspended sediment in the water.

The geology band combination is a neat application for finding geological features. This includes faults, lithology, and geological formations.

Figure 15: Agriculture

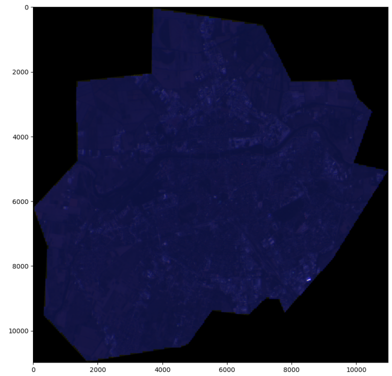


Figure 16: Bathymetric

Figure 17: Geology

Here again, for a better interpretation of the results, some more domain-specific knowledge is needed, so this remains an open issue and a point of improvement of this exploratory analysis.

4.2.2 Terrain model and Plovdiv elevation

Moving forward, the elevation profile of Plovdiv, and the bigger are of the tile where it is located is being examined.

Below, the three photos show the original elevation raster, the contours on it, as well as the histogram of elevation. Plovdiv is located in the Upper Thracian Plain, and as can be seen, is not very high, but it is surrounded by mountains on the South and on the North. There is still enough space for sunlight to warm the city, even though fog is also a common situation.

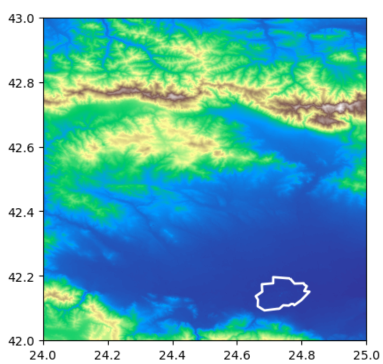


Figure 18: The whole tile with identification of where Plovdiv is

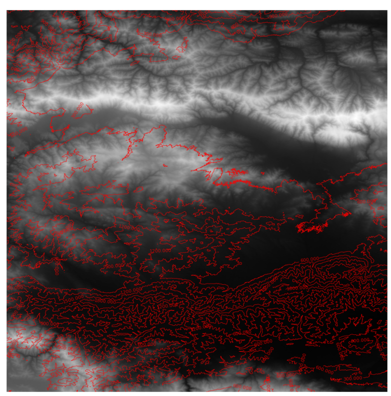


Figure 19: Contour lines

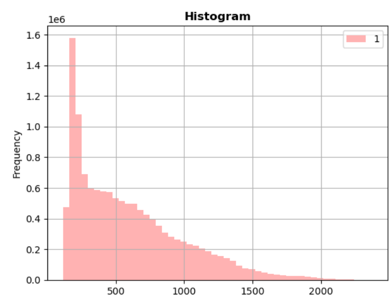


Figure 20: Altitude histogram

Next, as anticipated, the geometry of Plovdiv was used to mask the big layer with it, thus extracting the terrain model of the city. On it, the hills of Plovdiv can be seen clearly, as well as the river. From this diagram, it is evident that they are in the middle of the city, and as such, are also a main tourist attraction to reach easily and have a 360-degrees view of the city.

The statistics computed for the altitude according to these data are:

- Min: 130 m
- Max: 276 m
- Mean: 159 m

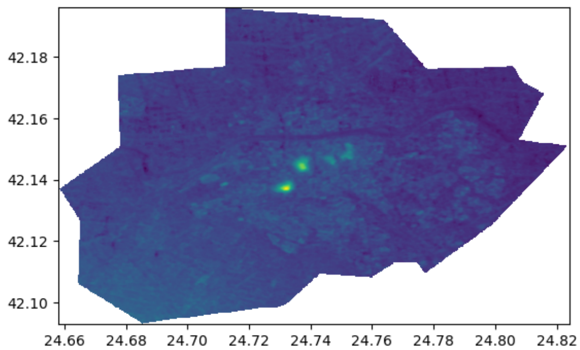


Figure 21: Terrain model Plovdiv

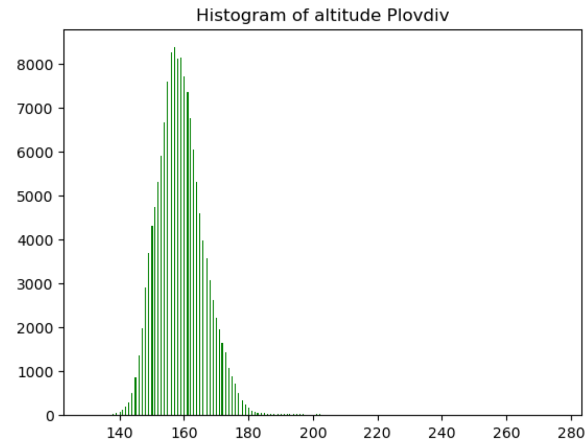


Figure 22: Altitude histogram

The hillshade of the city can also be examined, and it really appears that the hills have an effect of the shade on the city, also the river Maritza is recognizable, it again passing just a bit to the North from the city center.

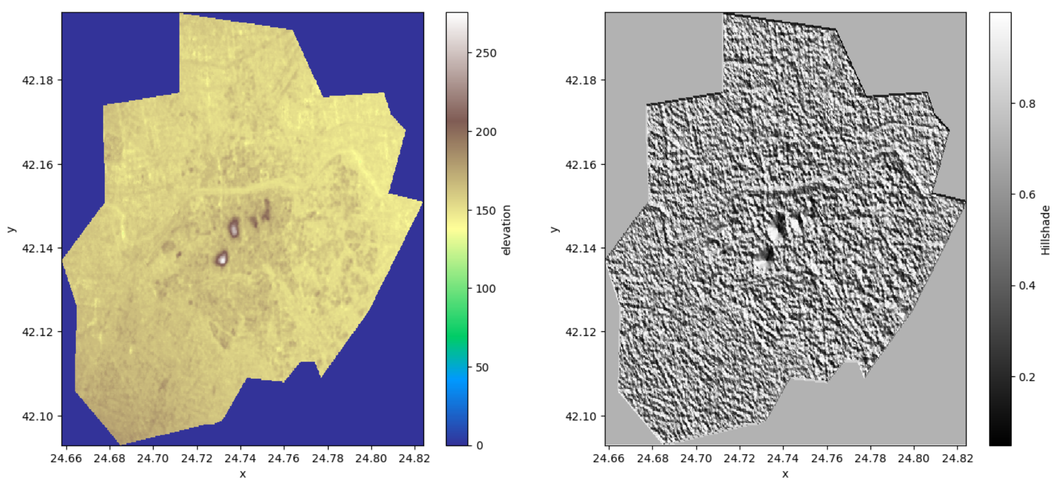


Figure 23: Hillshade plot

4.3 Open Steet Map analysis

After having dealt with the raster data, it was proceeded with analysis of of a different kind of information, useful to understand better the structure of the city. In particular, data from Open Street Map has been extracted, which contain resources on a number of different amenities in the city, as developed by contributors around the world. The protocol buffer around Plovdiv includes all that.

Following the previous analysis about the nature, the first thing to look at is the spatial structure of the city and what kinds of area are used for what purposes. Therefore, the landuse map is shown below. As a big city, the main residential area covers most of the center area, surrounded by farmland and industrial parts. On the barplot, one can compare what kind of usage the land of the city is utilized for. In fact, the distribution seems rather balanced in terms of usage for residence, industry, and agricultural activities, which is a nice trait for a city of the sort.

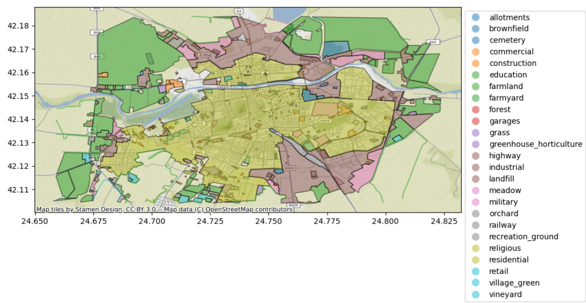


Figure 24: Landuse map

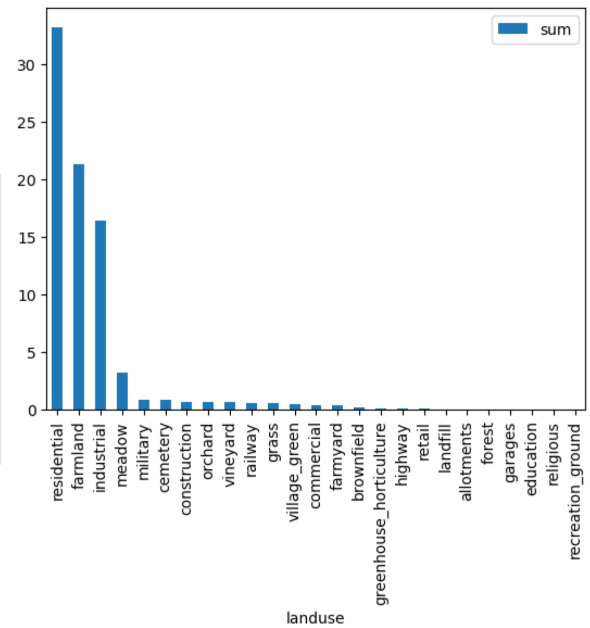


Figure 25: Landuse barplot

Then follows the buildings plot. It is coherent with the landuse one, as it shows the large concentration of residential buildings in the middle of the city. Another thing which is noted is that there are quite a bit hospitals in the western area of the city, it may therefore be thought of opening a new medical center (even a private one, or just a department) in the eastern parts to allow for a better and quicker help for persons in need. The diagram instead does not include residential and non-labeled entities, done with the purpose to analyze what kinds of services are the ones most available for citizens, and having a high number of schools is a good point as it allows for an efficient distribution of teaching personnel and children in the neighborhoods. However, there is a pretty high number of construction sites, and it is a trend from the last two years that many new residential buildings appear, making the urbanization quite a big problem for many issues such as traffic, parking posts, work, etc. In fact, these construction sites would change the structure and green and even farmlands are going to turn on to residential areas. Therefore, there will be a need to find a way to implement plans for green parks (even small ones) around the city.

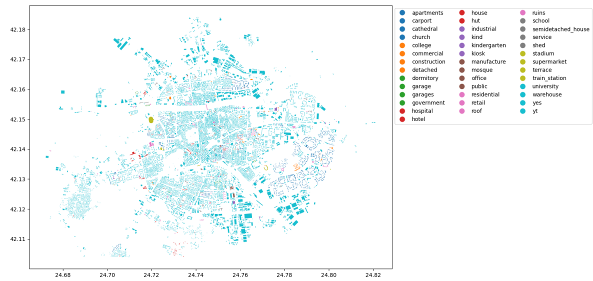


Figure 26: Buildings

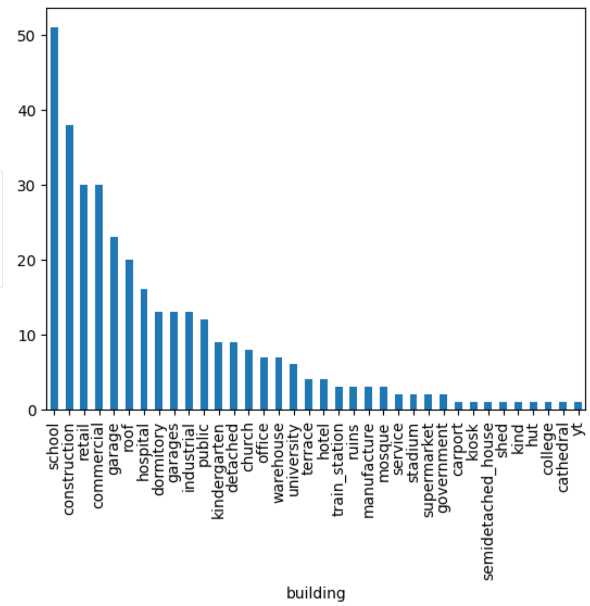


Figure 27: Buildings barplot

Next, in terms of networks, here below are shown the ones extracted from the data available containg walking paths, as well as the one containing the cycling roads. The city appears quite well connected even from remote parts, and there is a good potential for the usage of bikes, which until now has not yet been well proposed.

Figure 28: Walking

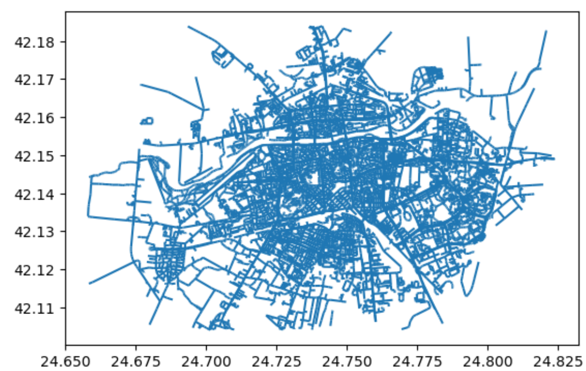


Figure 29: Cycling

When dealing with the terrain model of Plovdiv, the maximum and minimum points in terms of height were extracted with their coordinates. The lowest is on the wester part of the river Maritza, near the center for sport and relax The Rowing Channel, while the highest is on the top of one of the hills Plovdiv is famous for, in particular Dzhendem Tepe (or as is called by locals, The Youth Hill). By utilizing the network and functions, the shortest path between these two points is being shown here, just as an example of how a pedestrian could easily commute even from remote part of the city. Only the way up the hill proves to be ascending, and making curves walking up it.

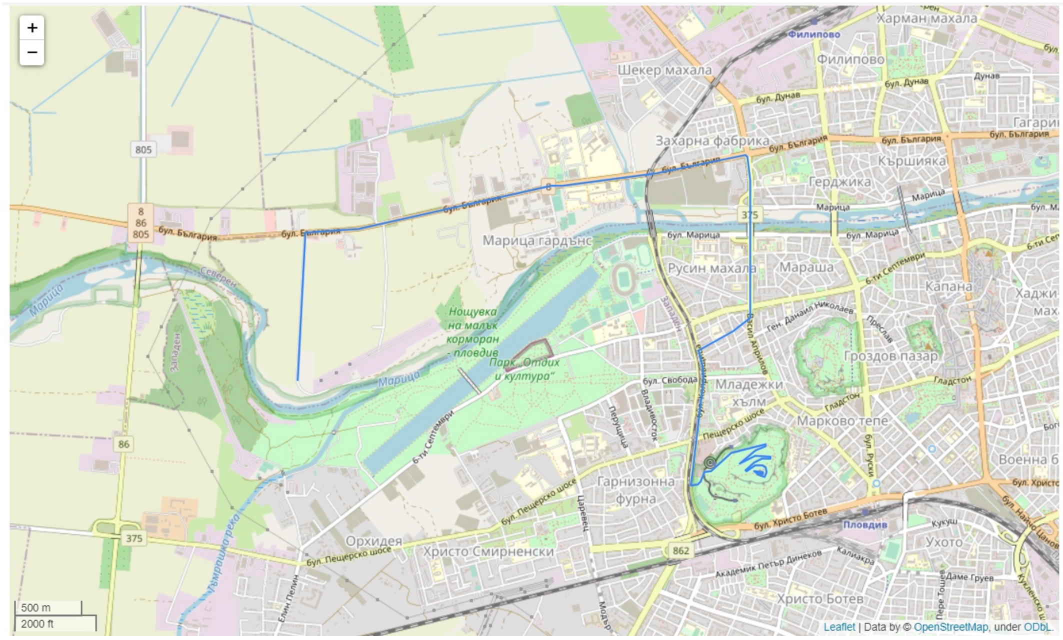


Figure 30: Shortest walking path from the lowest to the highest point of the city

5 Conclusions

To conclude the analysis, there are several issues to pint out. First of all, from the climate analysis, the most important things are the distribution of the variables in the regions, and the way this kind of information could be utilized for the purposes of sustainable development of the country of Bulgaria. By exploring how the climate is inside places, moreover some considerations regarding the energy consumption could be thought of in order to reduce it (depending on the temperature).

The satellite analysis is useful for the identification of places for a better city organization and checking of the health of the green areas. There is space to improve, and a good way to do it would be to create small parks inside the city, and a good way to take advantage would be to plant green areas in the round crossroads for example. The altitude and hillshade, on the other hand, are also to be considered for the purposes of deciding on what kinds of greenery would be able to grow effectively, by also considering the wind, temperature, and rain.

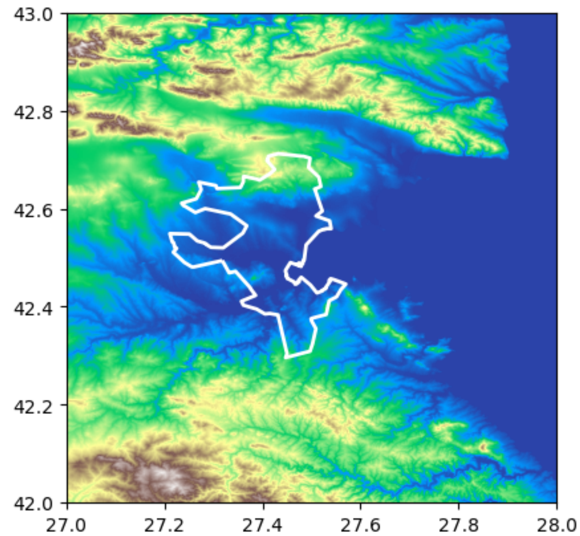
For the purpose of some regression analysis, as it has been seen, the spatial spillover effects in fact are not significant, but further exploration could be made by dividing into smaller regions, if such information could become available.

There is an ample space of improvement and further analysis after this initial exploration. By adding a time dimension, for example, it could be examined how the climate variables change and in particular whether some places have exhibited a change in them more prominent then others. The same could be searched for the NDVI index and check how it is developing with respect to the time dimension.

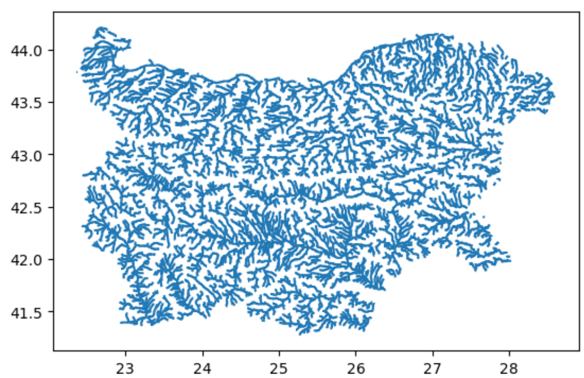
6 Appendix

In the project, there are some additional files that explore some other regions/subjects, but are discarded as not essential. However, they may be interesting to see.

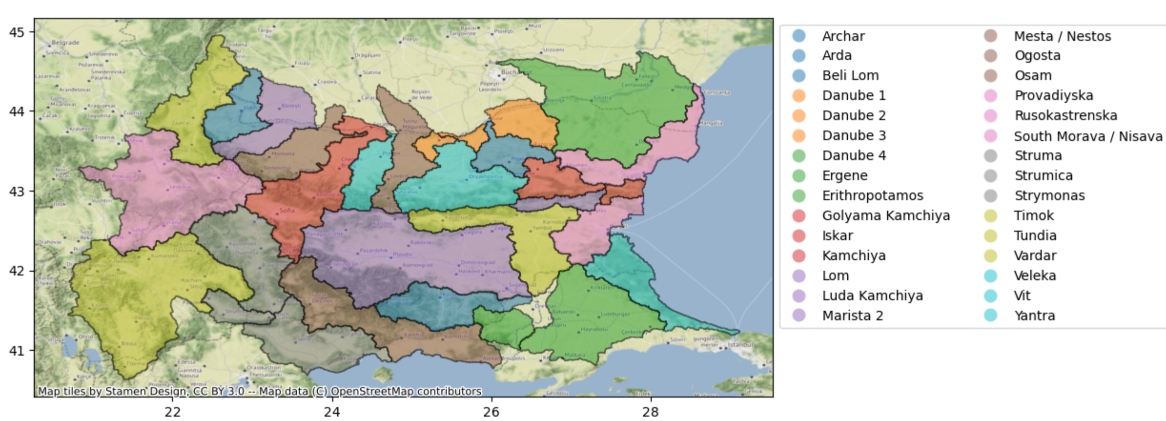
Those include first examination of the terrain and satellite bands for the city of Burgas, which is located on the Black Sea.



Also, a dataset about rivers was explored, showing the whole network clipped for Bulgaria, and had been done also clipping for the city. However, it was found that there some mistakes, so it was discarded.



Finally, a file which shows and explore the river basins and in particular has a map of the river basins located in Bulgaria. It was done with the intention to visualize some more characteristics combining with information from the National Statistical Institute (since some data about water usage, disposal etc. are divided according to river basins). After examining several geospatial datasets with basins of Europe and the world, there is not an exact match with how those are defined for the country of Bulgaria. Possibly this could be done at a later stage, by finding a way to match the geometries.



References

[1] National Statistical Institute. Home | national statistical institute. https://nsi.bg/en.

[2] WorldCLim. Historical climate data — WorldClim 1 documentation. https://worldclim.org/data/worldclim21.html.

[3] Copernicus open access hub. https://scihub.copernicus.eu/.

[4] The European Space Agency. Spatial - resolutions - sentinel-2 MSI - user guides - sentinel online - sentinel online. https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/resolutions/spatial.

[5] Earth data. https://www.earthdata.nasa.gov/.

[6] BBBike. BBBike extracts OpenStreetMap. https://extract.bbbike.org/.

[7] GISGeography. Sentinel 2 bands and combinations, 2022. https://gisgeography.com/sentinel-2-bands-combinations/.