1 Introduction

This is a document listing the methodological approach for the clustering of Athens urban habitats.

2 Methodology

The data used for the clustering is the output of the previous work on the Athens urban habitats. The data is a geojson file with 25 features, of which the following were used for the clustering analysis:

zipcode: the zipcode of the area

non_residential_density: the non-residential density of the area

non_residential_height: the average non-residential height of the area

residential_density: the residential density of the area

residential_height: the average residential height of the area

population_density: the population density of the area

green_area_percentage: the green area percentage of the area

imperviousness: the imperviousness of the area

vegetation: the vegetation of the area

roads: the road density of the area

water: the surface water of the area

trees_per_square_meter: the trees per square meter of the area

average_noise_level: the average noise level of the area

hot_days: the days with temperatures above 38 degrees Celcius in one summer in the area

mean_temperature: the mean temperature of the area

maximum_temperature: the maximum temperature of the area

mean_precipitation: the mean precipitation of the area

observations: the number of biodiversity citizen science observations of the area

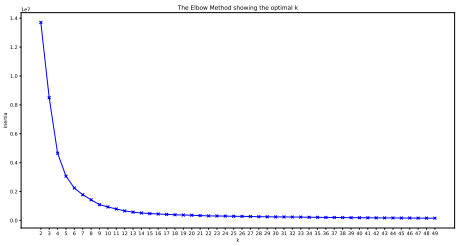

Figure 1: Elbow method for the clustering

And of course the label of the areas and their geometries.
The clustering was performed using the `scikit-learn` library in Python. The clustering was performed using the `KMeans` algorithm. In order to select the number of clusters, the `elbow` method was used. The `elbow` method is a method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset. The method consists of plotting the explained variance, or inertia, as a function of the number of clusters. The optimal number of clusters is where the inertia starts to decrease in a linear fashion. The inertia is defined as the sum of squared distances of samples to their closest cluster center. The inertia is calculated for each cluster and then summed up.

As can be seen in Figure 1, the optimal number of clusters is between 4 and 6. The clustering was performed for 4, 5 and 6 clusters and the silhouette score was calculated for each. The silhouette score is a metric used to evaluate the quality of a clustering. The silhouette score is defined for each sample and is composed of two scores:

a: The mean distance between a sample and all other points in the same class.

b: The mean distance between a sample and all other points in the next nearest cluster.

The silhouette score for a sample is then given by:

$$s = \frac{b - a}{max(a, b)} \tag{1}$$

The silhouette score for a set of samples is then given by the mean silhouette coefficient over all samples. The silhouette score ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. The silhouette score for the 4, 5 and 6 clusters is 0.31, 0.32 and 0.33 respectively. Since the silhouette score improvement is marginal, the clustering was performed for 4 clusters.

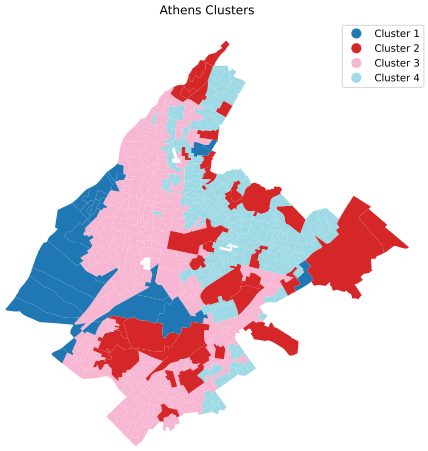

Figure 2: Clustering results

3 Results

The results of the clustering are shown in Figure 2.

In order to identify the clusters and their characteristics, the mean values of the features per cluster were calculated. The results are shown in Table 1.

As seen in Table 1, the clusters are characterized by different features. Cluster 1 is characterized by a high non-residential density and height, a low population density, low green area percentage and the highest amount of hot days. Cluster 2 is characterized by a very high percentage of green area, vegetation, a low amount of roads and water, a low non-residential density and height and low imperviousness.
Cluster 3 and 4 are characterized by a high residential density and height, a low non-residential density and height, a high population density. However Cluster 4 is to the east geographically, and is adjacent to more areas of Cluster 2, and as such more green area and vegetation, whereas Cluster 3 is to the west and is adjacent to Cluster 1, which has the highest number of hot days and non-residential density.
Finally for biodiversity, Cluster 2 is the most suitable, since it has the highest percentage of green area and vegetation, and has a critical amount of biodiversity observations. No other cluster has as many observations, while the number of observations of Cluster 1 can be attributed to its adjacency to Cluster 2 in the city center and the fact that it contains the Acropolis, which is a popular tourist attraction.

Finally, in Table 2 the ANOVA scores for the features are shown. The ANOVA scores are used to determine the importance of the features in the

Table 1: Mean values per cluster

cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Average Noise Level	69.430	67.326	69.645	67.909
Trees per square meter	0.003	0.003	0.004	0.004
Vegetation	0.108	0.222	0.083	0.092
Roads	8.785	5.173	9.782	3.054
Water	0.000	0.003	0.000	0.000
Residential Density	25.467	36.801	52.920	55.169
Residential Height	5.458	6.993	10.654	11.915
Non-Residential Density	30.305	3.927	0.555	0.404
Non-Residential Height	6.617	0.782	0.115	0.095
Population Density	0.014	0.010	0.030	0.034
Green Area Percentage	0.038	0.304	0.016	0.015
Imperviousness	79.266	52.235	85.756	87.080
Mean Precipitation	66.692	69.029	66.789	70.586
Maximum Temperature	44.845	44.632	44.584	44.144
Mean Temperature	29.708	29.167	29.604	29.088
Hot Days	13.154	12.086	11.316	10.317
Number of Observations	22.769	99.171	1.870	0.742

clustering. The ANOVA scores are calculated as follows:

$$\frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2} \tag{2}$$

where k is the number of clusters, n_i is the number of observations in cluster i, $\bar{x}_i$ is the mean of cluster i and $\bar{x}$ is the mean of all observations.

As can be seen in Table 2, the only feature that is not important in the clustering is the number of trees per square meter. The most important features are the non-residential density and height, followed by the maximum temperature, residential density, mean temperature, green area percentage and imperviousness.

Table 2: Sorted ANOVA scores for the clustered zipcodes

	ANOVA score	p value
Non-Residential Density	6.887E+02	2.805E-175
Non-Residential Height	6.784E+02	5.474E-174
Maximum Temperature	2.299E+02	4.415E-93
Residential Density	2.255E+02	6.776E-92
Mean Temperature	1.952E+02	2.857E-83
Green Area Percentage	1.675E+02	1.031E-74
Imperviousness	1.653E+02	5.125E-74
Vegetation	1.502E+02	4.891E-69
Mean Precipitation	1.495E+02	8.778E-69
Residential Height	1.097E+02	2.473E-54
Hot Days	8.938E+01	3.822E-46
Population Density	6.301E+01	1.814E-34
Roads	2.631E+01	8.593E-16
Average Noise Level	2.536E+01	2.887E-15
Number of Observations	2.299E+01	6.143E-14
Water	4.464E+00	4.176E-03
Trees per square meter	1.940E+00	1.222E-01