Exercise 8

PCA and HAC with GeoDa Solution

Part 1: Principal component analysis (PCA)

The PCA results show that the first three components explain **82.45% of the total variation** in the data, with **PC1 explaining 50.01%**, **PC2 explaining 23.32%**, and **PC3 explaining 9.12%**. This means that these three components capture most of the important patterns in the data, making it easier to understand.

The variable loadings are shown in Figure 1 below:

Variable Loadings:

	PC1	PC2	PC3
h0_19_15	0.296324	0.0148936	-0.214377
h20_39_15	0.299915	-0.0153345	-0.170605
h40_64_15	0.304247	0.0291103	-0.0992179
h65_79_15	0.268927	0.0934068	0.238838
h80plus_15	0.209998	0.107411	0.389878
f0_19_15	0.295026	0.0270326	-0.18739
f20_39_15	0.308317	0.00958395	-0.166043
f40_64_15	0.312309	0.037735	-0.0887993
f65_79_15	0.271282	0.100749	0.270013
f80plus_15	0.180028	0.0108916	0.438295
htot15	0.318843	0.0260373	-0.0944409
ftot15	0.320857	0.0385269	-0.0420552
lst_mean	-0.0130312	-0.217168	-0.379746
ndvi_mean	-0.0628626	0.247549	0.350684
no2_mean	0.068145	-0.427579	0.163457
pm10_mean	0.0781668	-0.411714	0.178322
pm25_mean	0.0824241	-0.411157	0.173318
bj_mean	0.0236953	-0.404106	0.0212954
bn_mean	0.0278428	-0.413073	0.059837

***Figure 1**: Variable loadings of the first three principal components (PCs).*

Note: The thematic maps below use natural breaks to classify data into 5 classes, ranging from negative to positive PC scores. Class 1 generally represents negative PC scores, while Classes 4 and 5 often reflect positive PC scores.

PC1 Results

PC1 shows **high positive loadings for all demographic variables**, including age groups for both males and females, as well as total male and female populations. For example, htot15 and ftot15 have loadings of +0.32. In contrast, environmental variables have low loadings on PC1, with the mean LST at -0.01 and the mean NDVI at -0.06.

This suggests that **PC1 represents areas with high overall population density**.

For example, on the thematic map (Figure 2), which uses 5 classes and natural breaks, areas with high PC1 scores (class 4 and 5, dark orange, 27 hectares) show an average population density of 492 people per hectare. In contrast, class 1 (yellow, 275 hectares) shows an average of 17 people per hectare.

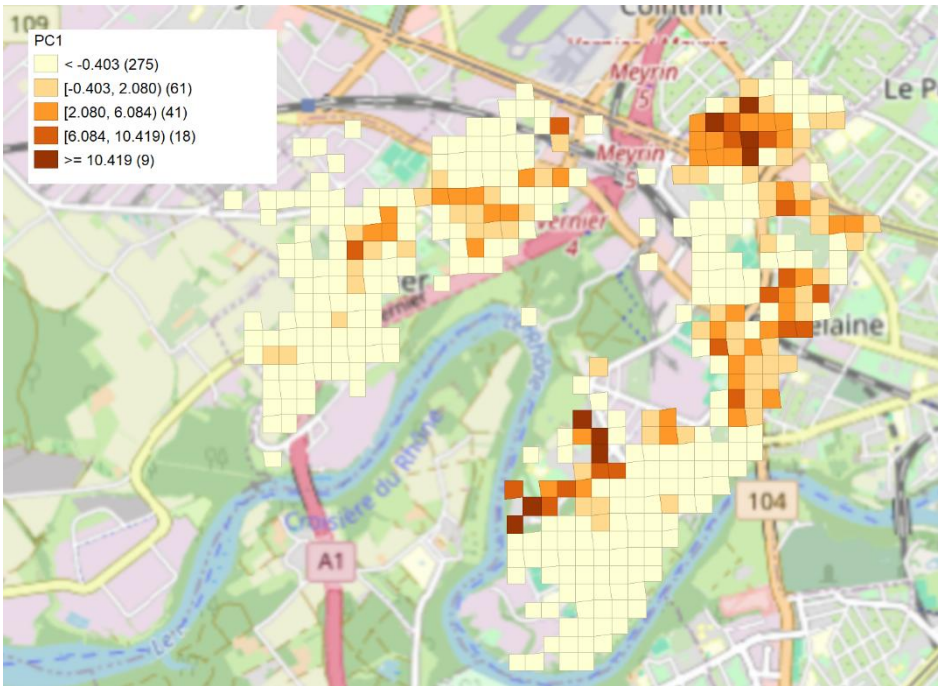


***Figure 2**: Thematic map of PC1. Orange and dark orange areas represent high positive PC1 scores, while yellow areas represent negative PC1 scores.*

PC2 Results

PC2 has high negative loadings for air pollution indicators and noise levels, with values such as NO_2 mean = -0.43, PM_{10} mean = -0.41, $PM_{2.5}$ mean = -0.41, noise levels around -0.4, and LST mean = -0.21. Conversely, NDVI shows high positive loading of +0.25, while demographic variables have low loadings on PC2.

This pattern suggests that **PC2 represents an environmental quality gradient**.

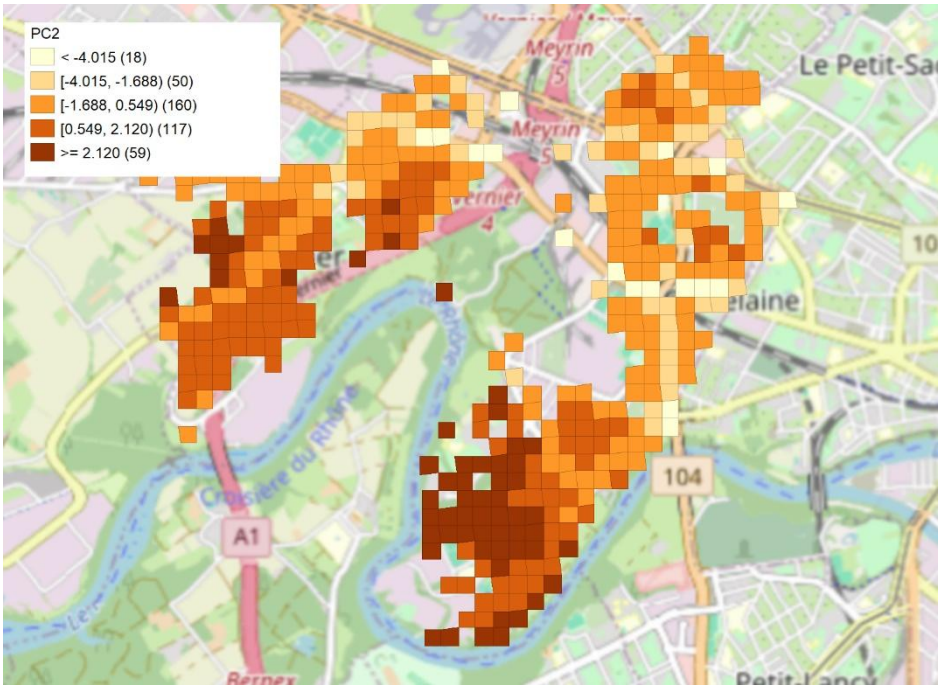


***Figure 3**: Thematic map of PC2. Yellow and light orange areas have negative PC2 scores, while dark orange areas have positive PC2 scores.*

For example, on the thematic map (Figure 3), category 5, which has high positive PC2 loadings and covers 59 hectares, shows a mean NDVI of 0.6, a mean $PM_{2.5}$ of 10.8 µg/m³, and a mean daytime noise level of 38.4 dB. In contrast, Categories 1-2, with negative loadings and covering 68 hectares, have a mean NDVI of 0.40, a mean $PM_{2.5}$ of 11.6 µg/m³, and a mean daytime noise level of 61.2 dB.

This indicates **that areas with higher PC2 scores generally have better environmental conditions**.

PC3 results:

PC3 shows high positive loadings for the elderly population (variables h80plus_15, f80plus_15, h65_79_15, and f65_79_15) and for NDVI (loading of +0.35). On the contrary, there are high negative loadings for the younger population (variables h0_19_15, f0_19_15, h20_39_15, and f20_39_15) and for LST (loading of -0.38).

This indicates that **PC3 reflects an age distribution gradient influenced by environmental factors**.

For example, on the thematic map (Figure 4), categories 4 and 5, representing areas with an older population (dark orange hectares), have a mean NDVI of 0.56, a mean LST of 28.8°C, a mean older population (65+) of 29.8, and a mean younger population (0-39) of 45.2. In contrast, Category 1, representing areas with a younger population, has a mean NDVI of 0.33, a mean LST of 32.4°C, a mean older population (65+) of 13.0 and a mean younger population of 151.5.

These results suggest that **areas with older residents tend to be greener and cooler.**

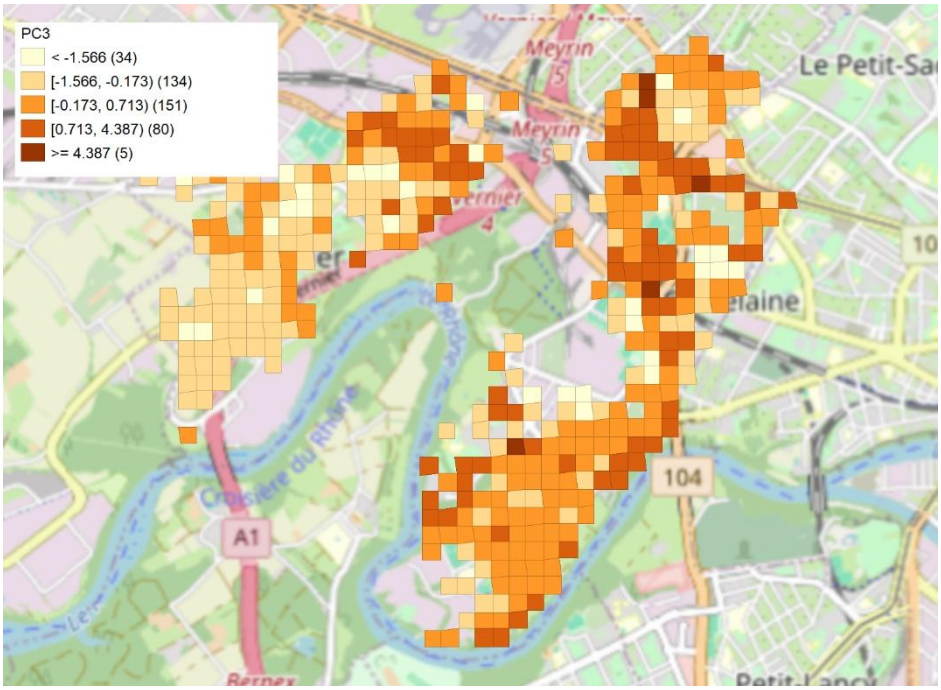
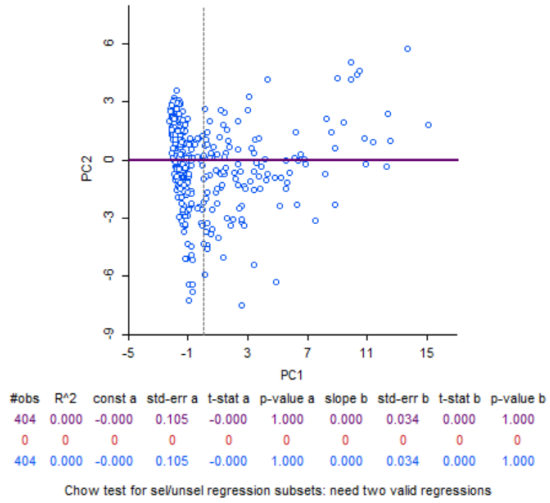


***Figure 4** : Thematic map of PC3. Yellow and light orange areas have negative PC2 scores, while dark orange areas have positive PC2 scores.*

Scatterplot of PC1 vs. PC2

Figure 5 shows the scatter plot of PC1 versus PC2, illustrating the relationship between population density and environmental quality. The plot can be interpreted by quadrants:



- **Top right**: High population density, good environmental quality.
- **Bottom right**: High population density, poor environmental quality.
- **Top left**: Low population density, good environmental quality.
- **Bottom left**: Low population density, poor environmental quality.

***Figure 5**: Scatterplot of PC1 vs PC2*

Part 2: Hierarchical Clustering (HAC)

Figure 6 shows the results of hierarchical clustering (HAC) with five distinct classes. Each cluster is represented by a unique color: cluster 1 in light blue, cluster 2 in blue, cluster 3 in light green, cluster 4 in dark green, and cluster 5 in pink.

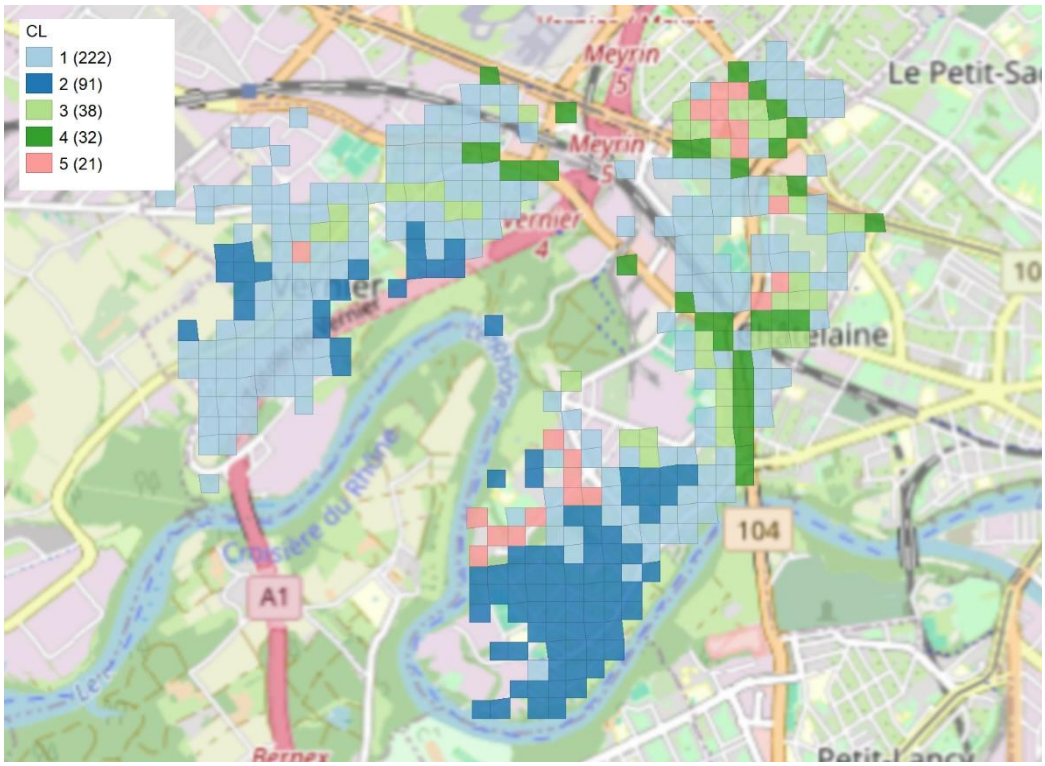


***Figure 6**: HAC clustering using 5 classes, with each cluster color-coded: Cluster 1 in light blue, Cluster 2 in blue, Cluster 3 in light green, Cluster 4 in dark green, and Cluster 5 in pink.*

Cluster profiles

To characterize each cluster, the mean values of key demographic and environmental variables, along with the principal components (PC1, PC2, and PC3), were calculated for each cluster. This approach enabled the identification of patterns related to population density, environmental quality, and age distribution, providing a comprehensive basis for interpreting each cluster's unique profile (Table below).

CL	total_population	young_population	old_population	ndvi_mean	lst_mean	no2_mean	pm10_mean	pm25_mean	bj_mean	bn_mean	PC1	PC2	PC3
1.0	42.13	21.40	6.02	0.49	30.77	247.26	161.40	109.29	52.20	42.15	-0.99	-0.23	-0.08
2.0	15.18	6.37	3.16	0.57	28.87	225.06	156.26	106.60	40.19	30.57	-1.82	2.03	0.15
3.0	276.87	157.80	26.34	0.40	31.05	261.25	166.12	111.98	53.26	43.50	4.15	-0.80	-1.25
4.0	87.93	50.12	11.03	0.39	31.34	323.50	179.66	119.35	63.45	54.75	0.67	-4.56	0.75
5.0	502.94	235.28	102.06	0.49	29.19	244.78	161.97	109.84	46.65	37.10	9.78	2.03	1.34

***Table 1**: Mean characteristics of demographic, environmental variables, and principal components (PC1, PC2, and PC3) for each cluster.*

Cluster 1 (light blue, 222 hectares) **includes moderately populated areas with average pollution levels**.

The principal component scores indicate slightly below-average population density (PC1: -0.99), moderate environmental quality (PC2: -0.23), and a balanced age distribution (PC3: -0.08). These areas are likely to have low population density, balanced age demographics, and some green spaces.

Cluster 2 (blue, 91 hectares) **represents areas with low population density, high vegetation, and low pollution**.

With the lowest population density (**PC1: -1.82**), high environmental quality (**PC2: +2.03**), and a slightly older demographic (PC3: +0.15), these areas are characterized by more vegetation, cleaner air, and a somewhat older population.

Cluster 3 (light green, 38 hectares) **includes highly populated areas with moderate pollution levels and younger population**.

It has high population density (**PC1: +4.15**), lower environmental quality (PC2: -0.80), and a younger population (**PC3: -1.25**), suggesting densely populated zones with less green space, younger residents, and higher pollution and noise levels.

Cluster 4 (green, 32 hectares) **represents areas with moderate population density and** slightly older populations **but severe pollution**.

This cluster scores moderately on population density (PC1: +0.67), has the lowest environmental quality (**PC2: -4.56**), and has a slightly older population (PC3: +0.75). These areas are high-traffic zones with significant environmental degradation despite moderate population density.

Cluster 5 (pink, 21 hectares) **features areas with very high population density, good environmental quality and older demographics.**

With very high population density (PC1: +9.78), good environmental quality (PC2: +2.03), and an older population (PC3: +1.34), these areas are densely populated, have significant elderly populations, and maintain better environmental conditions than Cluster 4.