Factor Analysis for Athens, Greece

Konstantinos Ziliaskopoulos

March 24, 2023

1 Introduction

The purpose of this analysis is to perform a factor analysis on the data of Athens, Greece. Factor analysis has been used as a tool for analysing extreme heat vulnerability in cities, [1]. The data is from the 2011 census and contains socio-economic data for each zipcode area in Athens. The goal is to find the underlying factors that explain the variance of the data, and identify the most important variables that contribute to each factor. The analysis is performed in python.

2 Data

The data used in this analysis are from the 2011 census of Athens, Greece on a zipcode level. The processed data is available on the github repository of the project, as a geojson file.

3 Methodology

The data was initially merged together from different excel files and joined with the shapefile of the zipcodes. The data was then spatially joined with the case study region in Athens in order to isolate the zipcodes of the case study region. The final version of the dataset had data for 726 zipcodes, with 60 variables for each zipcode area. Since we will perform a factor analysis, which is a form of oblique principal component analysis, we will need to scale the data. The data is therefore scaled and standardized to have a mean of 0 and a standard deviation of 1, with the following formula:

$$\frac{x - \mu}{\sigma} \tag{1}$$

where x is the value of the variable, μ is the mean of the variable and σ is the standard deviation of the variable.

After scaling, two statistical tests were performed to identify whether the data is suitable for a factor analysis. The first test is the Bartlett's test of sphericity. This test is used in factor analysis to determine whether the correlation matrix of the observed variables is suitable for factor analysis. The test assesses the hypothesis that the correlation matrix is an identity matrix (i.e., all diagonal elements are 1 and all off-diagonal elements are 0), which indicates that there are no correlations among the variables that cannot be accounted for by a set of factors. If the test statistic is significant (i.e., the p-value is less than the significance level), then the null hypothesis is rejected, indicating that the correlation matrix is not an identity matrix and that factor analysis may be appropriate.

The second test is the Kaiser-Meyer-Olkin measure of sampling adequacy. This measure is used in factor analysis to assess whether the data are suitable for factor analysis. The KMO measure ranges from 0 to 1, with values closer to 1 indicating that the data are more suitable for factor analysis. A KMO value of 0.6 or higher is generally considered acceptable, while values below 0.5 indicate that the data are not suitable for factor analysis. The KMO measure assesses the degree of common variance among the variables, and values closer to 1 indicate that there is a high degree of common variance, which is necessary for factor analysis. The results of the tests are shown in Table 1 and Table 2.

Table 1: Bartlett Test

chi-square	283445.417551
p-value	0.000000

Table 2: Kaiser-Meyer-Olkin Measure of Sampling Adequacy

Kaiser-Meyer-Olkin (KMO)	0.8178

According to the results of the tests, the dataset is a good candidate for a factor analysis. In order to identify the proper amount of factors to use, we will use the Kaiser criterion and the scree plot. The Kaiser criterion states that factors should only be retained if their eigenvalues are greater than 1. The scree plot is a plot of the eigenvalues of the factors, which can be used to visually identify the number of factors to retain. The results of the Kaiser criterion and the scree plot are shown in Figure 3. While the results of Figure 1 indicate that at most 10 factors should be retained, the scree plot indicates that most of the variance is explained by the first 5 factors. As such, we will retain 5 factors in the final analysis. As for the rotation method, we will use the varimax rotation method, which is the most commonly used rotation method in factor analysis and is recommended for exploratory factor analysis for components with a high degree of correlation.

4 Results

In Figure 4, we can see the loading coefficients of the original features in each of the five factors. The loading coefficients are the correlation between the original features and the factors. Only loading coefficients with an absolute value greater than 0.4 are shown.

As shown in Figure 4, each factor constructs a different citizen profile. Factor 1 suggests a citizen profile of an individual who is between 14-44 years old, is unemployed and currently not enrolled in school, is an immigrant from Eastern Europe or a developing country from Asia, the Middle East or North Africa, is socially deprived and rents relatively small, old apartments.

Factor 2 suggests a citizen profile of an individual who is retired, predominately over 65, is a Greek national or from a developed country in Europe and rents more often than not.

Factor 3 suggests a citizen profile of an individual who is between 0-64 years old, is employed or studying, is a Greek national, lives in a relatively large, new apartment and is a home owner.

Factor 4 suggests a citizen profile who is financially independant or subsides from investments, is either a dual citizen or from a developed country outside of Europe and lives in very large

apartments or houses.

Factor 5 suggests a citizen profile of an individual who is unemployed or has irregular employment, is from a developing country in Asia, the Middle East or North Africa and has an irregular housing situation, such as communal housing.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
SS Loadings	11.271153	8.551265	8.499131	5.140025	4.846377
Proportion Var	0.187853	0.142521	0.141652	0.085667	0.080773
Cumulative Var	0.187853	0.330374	0.472026	0.557693	0.638466

Table 3: Factor loadings and proportion of variance explained by each factor.

In Table 3, we can see the proportion of variance explained by each factor. The first factor explains 18.8% of the variance, the second factor explains 14.3% of the variance, the third factor explains 14.2% of the variance, the fourth factor explains 8.6% of the variance and the fifth factor explains 8.1% of the variance. As such, the first five factors explain 63.8% of the variance in the dataset.

References

[1] Seema G Nayak, Srishti Shrestha, PL Kinney, Zev Ross, SC Sheridan, CI Pantea, WH Hsu, Nicola Muscatiello, and Syni-An Hwang. Development of a heat vulnerability index for new york state. *Public Health*, 161:127–137, 2018.

Figure 1: Kaiser Criterion

Index	Eigenvalues
1	17.65
2	8.94
3	5.76
4	4.56
5	2.68
6	1.56
7	1.46
8	1.32
9	1.21
10	1.10
11	0.98
12	0.90
13	0.89
14	0.85
15	0.83
16	0.76
17	0.70
18	0.66
19	0.57

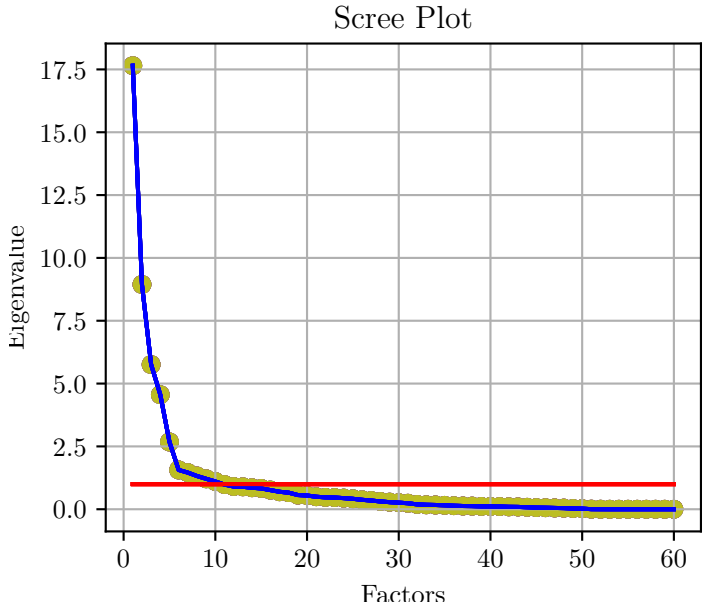

Figure 2: Scree plot, created with `matplotlib` in python. The red line indicates the Kaiser criterion threshold of 1.

Figure 3: Kaiser criterion and scree plot

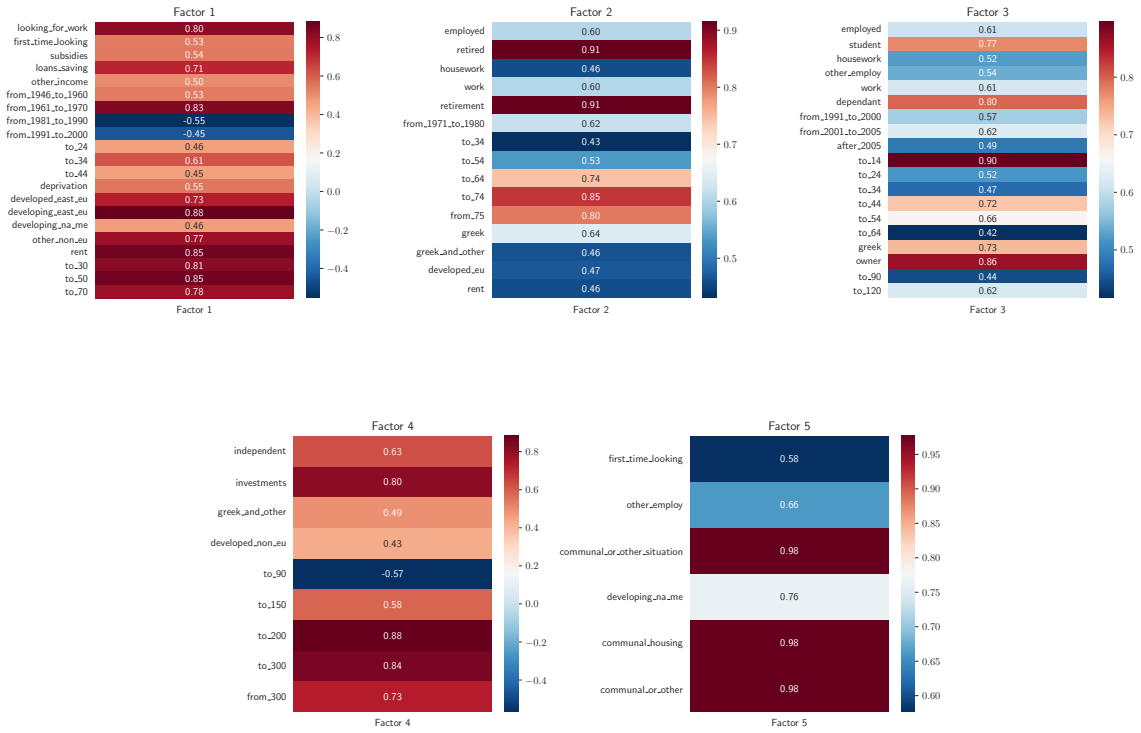

Figure 4: Heatmaps for the loading coefficients of the original features in each of the five factors from the factor analysis. Created with `matplotlib` in python.

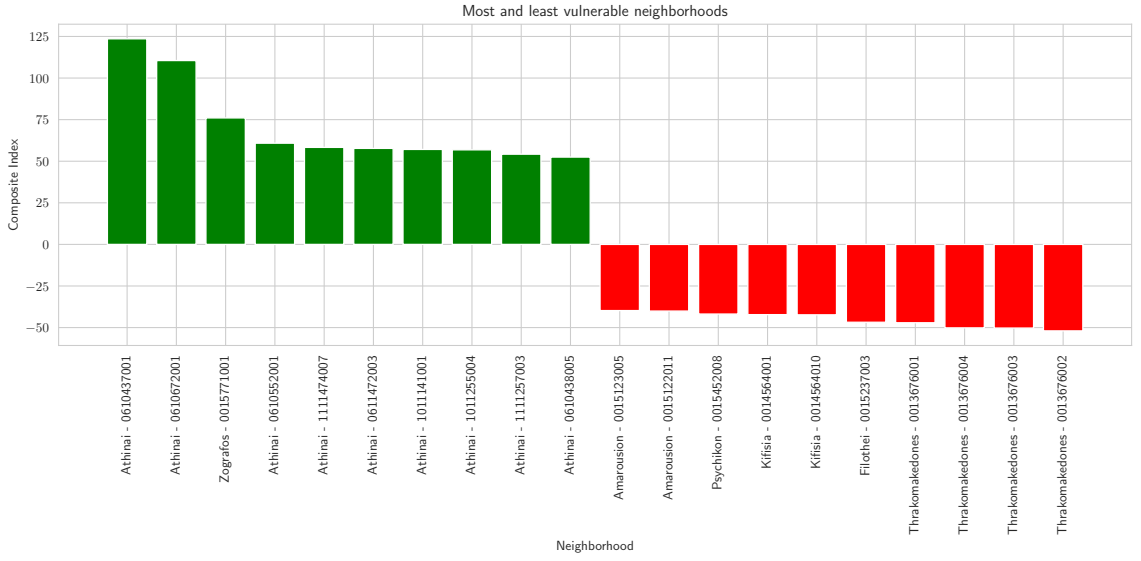

Figure 5: Top 10 and bottom 10 municipalities in terms of the composite index.