Analyzing the determinants of crime in the city of Helsinki

Project for the course MS-E2112

Mikko Kaivola 525789

12th March 2021

1 Research questions

During last summer, I explored how violent crime and property crime were associated with different social, demographical and environmental covariates in the city of Helsinki. The densities of bars and different transportation stations were crucial in accounting the observed variation in the crime data when using Poisson regression models. Unfortunately, many crucial covariates identified from the literature had to be excluded due to strong multicollinearity affecting the stability of the regression coefficients. The distribution of bars was also concentrated around the city center where some of the highest crime counts were observed. Thus it is quite likely that the effect size of bar density was mainly driven by these locations.

This project aims to analyze these excluded covariates by first performing a dimensionality reduction using principal components and then estimating their associations with the crime counts. PCA is also used to explore the structure of the covariate data to see whether it is possible to identify components that account for large amounts of variation and whether these components can be interpreted in a meaningful fashion.

2 Data

The crime data was requested from the Helsinki police department while the non-spatial data was collected from the Aluesarjat-database. The density variables were formed by estimating probability density functions via kernel density over the Helsinki region using the location of bars and stations as observations. A more thorough explanation of the variables is shown in Table 1 and in my bachelor's thesis. The number following the name describes the year the variable values were observed. In contrast to the bachelor's thesis, only those areas for which the crime counts and other variables were known exactly were used in this analysis.

Table 1: Variable names and their explanations.

Variable name	Explanation
People2017	Number of people living in an area
Work2015	Number of jobs in an area
AEDensity	Density of establishments providing alcohol
Africa2017	Number of people of African descent
Asia2017	Number of people of Asian descent
PerusA2017	Number of adults having only a primary school education
YH2017	Number of single parents
VierK2017	Number of people having a non-native first language
GINI2015	Gini coefficient between households
TBSTDen	Density of bus and train stations
OmaisR2017	Number of property crimes
HenkiR2017	Number of violent crimes

Univariate analysis

Since the number of variables is relatively large, a criterion was needed to reduce the scale of the uni- and bivariate analyses. Interesting variables were identified by their departures from normality measured using the sample skewnesss and excess kurtosis coefficients. The descriptive statistics and their confidence intervals are shown in Table 2.

Table 2: Descriptive statistics.

	Mean	SD	Median	MAD	Min	Max	Skewness	Kurtosis
Explanatory variables								
People2017	6628.63	4260.37	6007.00	3211.31	408.00	15487.00	0.42	-0.75
	(5377.69 , 7965.19)	(3366.35 , 4915.67)	(5050 , 7414)	(1879.94 , 6650.94)			(0.03 , 0.83)	(-1.32 , 0.4)
Work2015	4370.61	6358.83	1711.00	1930.35	0.00	25953.00	2.13	3.89
	(2597.18 , 6471.35)	(3298.64 , 8435.44)	(925 , 3080)	(982.96 , 3836.97)			(1.24 , 3.22)	(0.12 , 11.61)
AEDensity	24088.69	48004.54	10514.53	10994.63	0.00	256539.56	3.48	12.61
	(11799.8 , 40577.53)	(14884 , 72790.97)	(4236.43 , 11786.99)	(6044.77 , 16356.01)			(1.51 , 4.94)	(1.28 , 25.86)
Africa2017	202.32	223.04	126.00	170.50	0.00	973.00	1.37	1.72
	(139 , 274.19)	(153.75 , 287.28)	(46 , 246)	(66.72 , 268.35)			(0.44 , 2.05)	(-1.23 , 4.73)
Asia2017	301.51	305.63	231.00	244.63	0.00	1521.00	1.69	3.91
	(213.13 , 398.85)	(200.16 , 409.65)	(97 , 351)	(127.5 , 370.65)			(0.3 , 2.38)	(-1.29 , 7.38)
PerusA2017	756.93	554.26	735.00	527.81	16.00	2369.00	0.85	0.45
	(593.49 , 931.04)	(408.3 , 677.58)	(494 , 977)	(339.52 , 695.34)			(0.14 , 1.33)	(-0.95 , 2.22)
YH2017	272.00	186.08	256.00	161.60	14.00	825.00	0.87	0.73
	(218.29 , 330.88)	(133.46 , 230.93)	(197 , 327)	(105.26 , 226.84)			(0 , 1.35)	(-0.94 , 2.61)
VierK2017	1025.59	931.81	863.00	864.36	5.00	4347.00	1.33	2.10
	(760.66 , 1329.26)	(635.32 , 1205.79)	(444 , 1255)	(508.53 , 1165.32)			(0.26 , 1.93)	(-1.14 , 4.88)
GINI2015	0.31	0.11	0.26	0.04	0.19	0.74	2.09	4.74
	(0.28 , 0.34)	(0.06 , 0.15)	(0.25 , 0.29)	(0.02 , 0.08)			(0.92 , 3.04)	(-0.38 , 10.78)
TBSTDen	26410.60	45227.63	290.04	430.02	0.00	242706.55	3.03	10.96
	(14627.5 , 42049.15)	(20653.47 , 67559.05)	(0 , 39106.44)	(0 , 45183.57)			(0.26 , 3.88)	(-1.87 , 18.15)
Dependent variables								
OmaisR2017	527.17	663.18	310.00	308.38	3.00	3290.00	2.68	7.46
	(350.37 , 748.85)	(283.35 , 943.55)	(230 , 497)	(177.91 , 450.71)			(0.86 , 3.81)	(0.09 , 17.52)
HenkiR2017	89.98	120.14	58.00	51.89	1.00	568.00	2.51	6.17
	(57.2 , 131.36)	(53.75 , 167.58)	(35 , 71)	(28.17 , 75.61)			(1.23 , 3.68)	(0.57 , 16.01)

Sample size n = 41, 95 percent percentile bootstrap confidence intervals shown in parenthesis

From Table 2, it is quite clear that the density-based variables have the most non-normal distributions using the given measure. They also exhibit the greatest differences between non-robust and robust measures of location and scatter. This behavior is mostly driven by two observational units as can be seen in Figure 1. These units are also outlying in terms of crime counts as shown in Figure 2 which probably explains why the count variables depart quite strongly from normality.

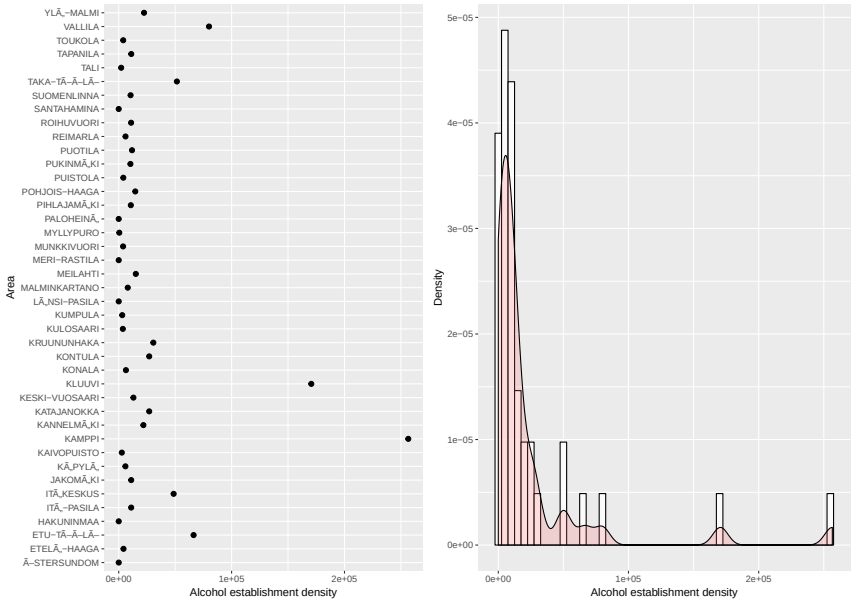

Figure 1: Dot plot and histogram for alcohol establishment density.

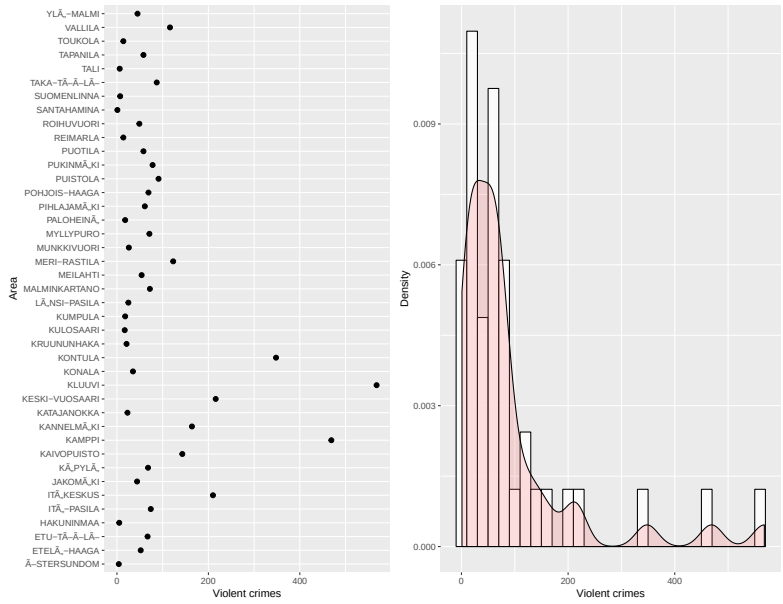

Figure 2: Dot plot and histogram for violent crime.

Bivariate analysis

Many covariates had to be excluded from the original regression analysis. The reason for this can be seen from the Pearson correlation coefficient matrix with bootstrapped 95 percent confidence intervals shown in Figure 3. There seems to be very high correlations between low education level, single parenthood and the number of people of different foreign backgrounds. In contrast, the income inequality covariate is negatively correlated with all aforementioned variables. It also seems that the total population size correlates with different subpopulation sizes. A similar effect is observed between alcohol establishment density and the number of jobs. The most anomalous behavior is seen in the confidence intervals corresponding to station density, number of jobs and alcohol establishment density which are quite wide and highly asymmetric. This is partly explained by Figure 1 where most of the observed densities are clustered around specific values. That aside, these correlations give hint to a possible structure in the explanatory variables which is analyzed in Section 3.

To assess the associations, Pearson correlations between the response and explanatory variables are shown in Figure 4. The values are a bit misleading since the outlying observational units mask some of the relevant dependencies which might actually be non-linear. An example of this is shown in Figure 5 for the single parent covariate. This kind of behavior was one of the reasons why the Poisson regression model was used in the bachelor's thesis and should be noted when analyzing the results of any linear method. One could consider non-linear transformations to linearize the associations but this would most likely make the PC transformations even more difficult to interpret. Thus all covariates retain their original scales.

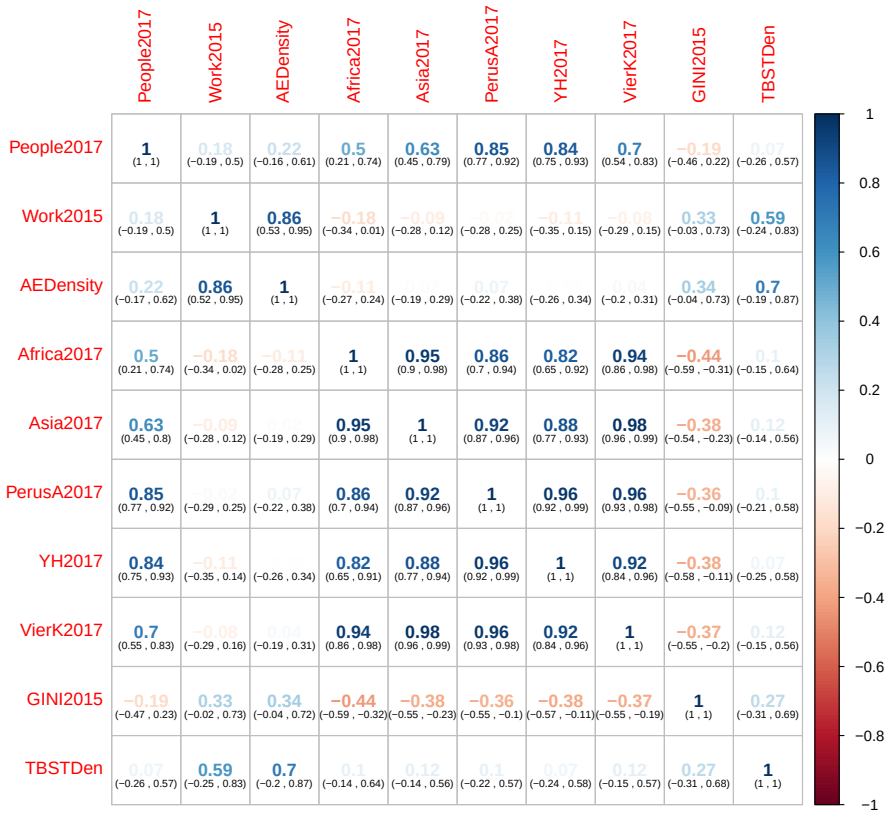

Figure 3: Pearson correlation coefficient matrix for the explanatory variables.

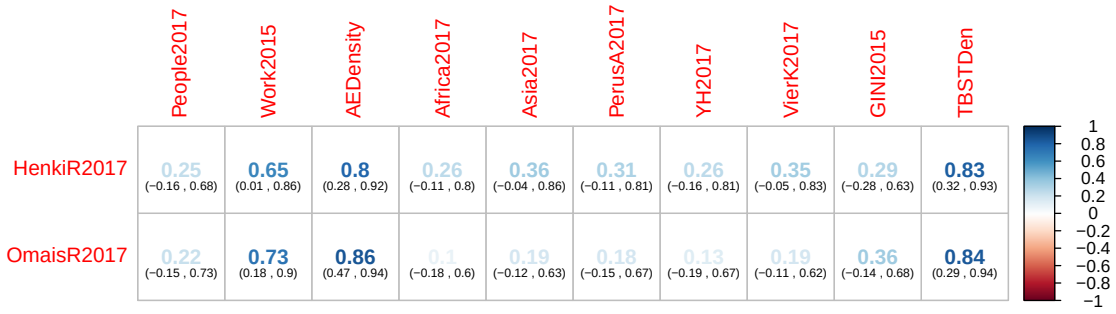

Figure 4: Pearson correlation coefficients between response and explanatory variables.

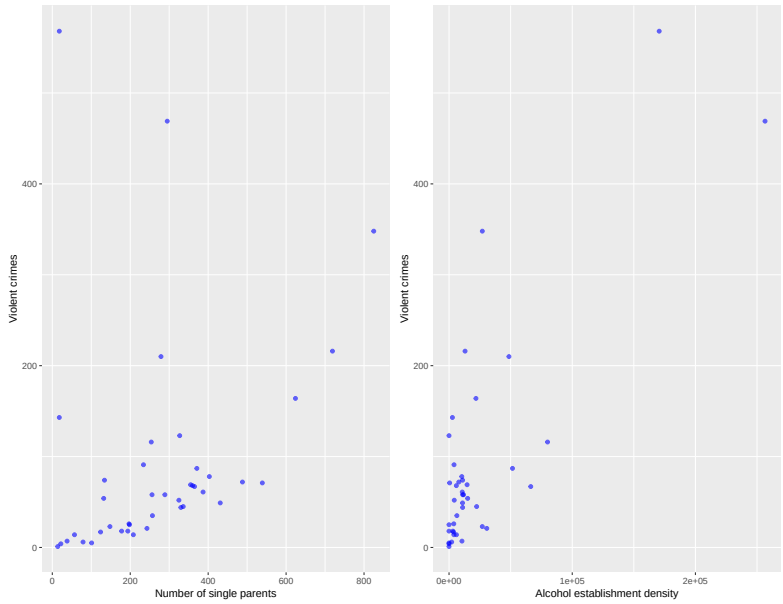

Figure 5: Scatter plots demonstrating different types of dependencies in the data.

3 Multivariate analysis

Since the scales of the variables differ greatly, correlation based PCA was performed to prevent the density-based variables from dominating the total variance. The variance statistics of the principal components are shown in Table 3 and the correlations between them and the original covariates in Figure 6. The proportions of variance seem to behave quite nicely with "geometric" proportion decay for the first three components.

Table 3: Variance statistics of the principal components.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	2.34	1.63	0.85	0.81	0.55	0.36	0.20	0.16	0.13	0.08
Proportion of variance	0.55	0.27	0.07	0.06	0.03	0.01	0.00	0.00	0.00	0.00
Cumulative proportion	0.55	0.81	0.88	0.95	0.98	0.99	1.00	1.00	1.00	1.00

Not surprisingly, the highly correlated demographical and social variables from Figure 3 load strongly on the first component with Gini having an opposite signed loading whereas the density-based covariates and the number of jobs load on the second component. The third component is much harder to analyze since no covariate loads to it strongly. Gini is quite spread out between the first four components with the largest absolute loading on the fourth component. The reasons for some of these loadings become clearer when examining the observational units having the highest contributions to a given component. Only the first four components are analyzed in detail since they account for 95 percent of the total variance and the correlations are rather small in magnitude for the remaining components. The ten observational units having the highest contributions for a given component are shown in Figure 7 with the lines indicating what the contributions would look like if they were uniformly distributed among the units.

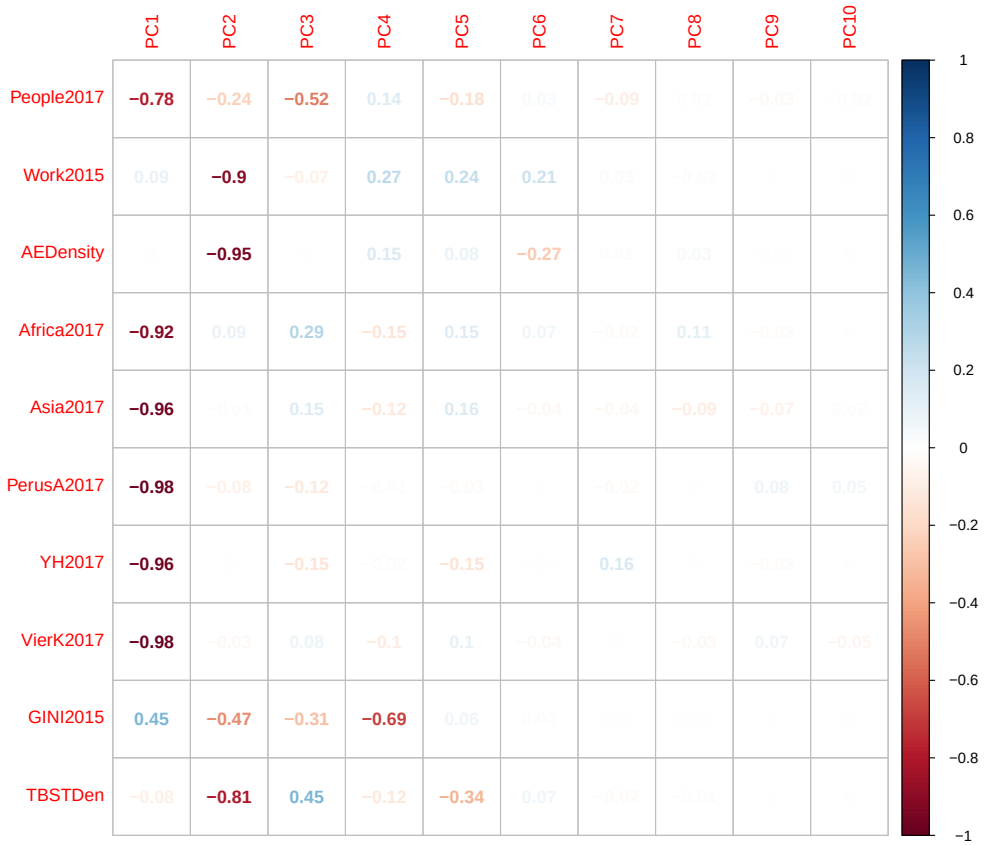

Figure 6: Correlations between the original covariates and the principal components.

Looking at the contributions in general, it is possible to identify two types of principal components: those in which only a few observational units constitute the majority of the variance and those in which the contributions are more spread out among the units. PC2 and PC4 are the former while PC1 and PC3 are the latter. The two dominating units in PC2 are located at the city center where the bar and station density is very high and approximately quarter of the jobs are located in the data set. PC4 exhibits a similar phenomenon where two areas of relatively high income inequality contribute greatly to the variance. Almost all of the remaining areas in PC4 seem to belong to the opposite ends of the income inequality values. This sort of phenomenon characterizes the contributions of PC1 and PC3 which can be seen quite clearly from Figure 8.

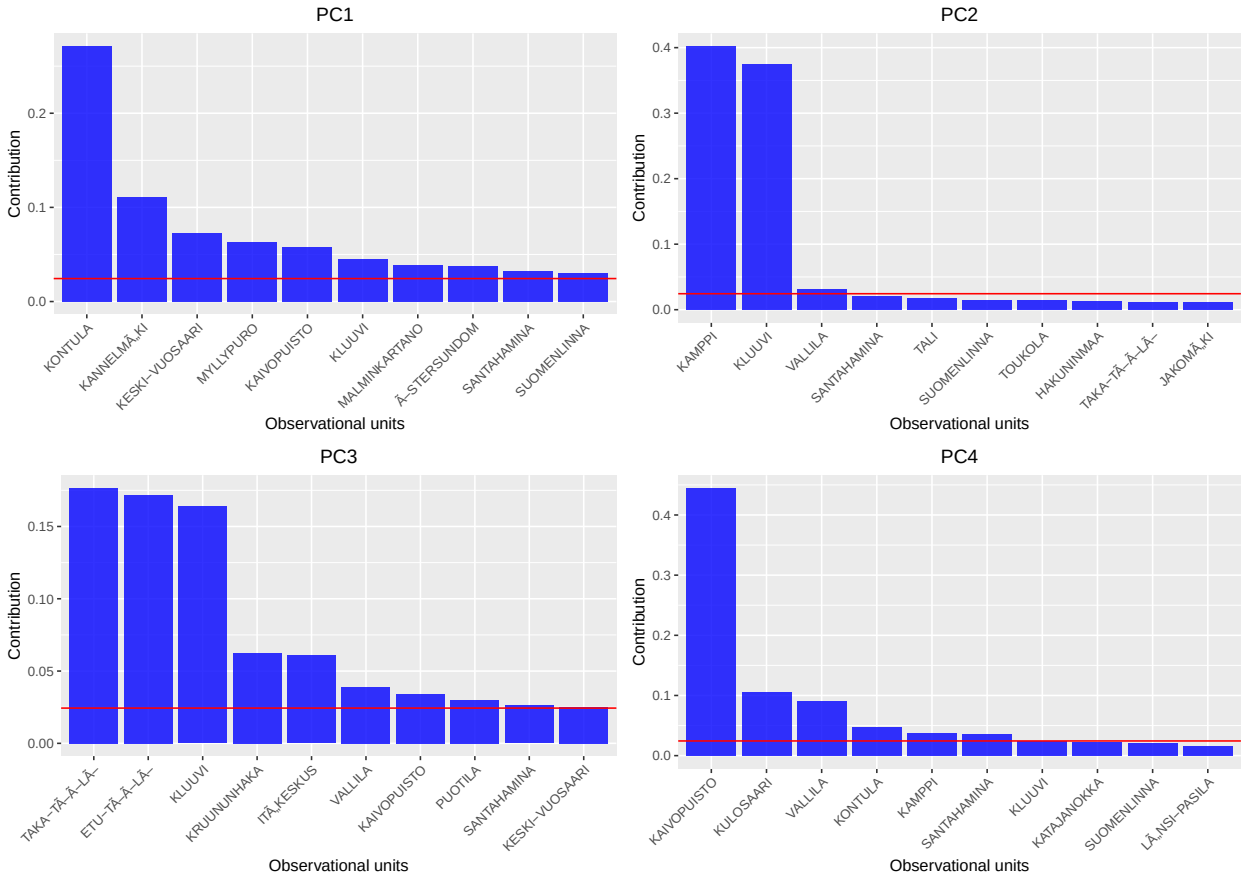

Figure 7: Contributions of observational units to different principal components

The PC1 in Figure 8 seems to describe areas in terms of their social and demographical character. Areas to the left tend to have more people from populations which are associated with higher crime rates while the areas on the right tend to the opposite direction. The main contributors to the variance of the component are the units with extreme values on both sides of the spectrum. PC3 seems to divide areas in terms of the source of people flow: in Kluuvi, most of the traffic is caused by commuters coming to work or going home while in Etu- and Taka-Töölö most flow is caused by the people living there rather than those commuting. This sort of characterization could be useful since one of the problems concerning the modelling of crime in the bachelor's thesis was the activity space of the criminals. It was not possible to take into account whether the crime was done by a person living in the given area or by an outsider. However, the relatively low contribution of the component to the total variance and the magnitude of the correlations in Figure 6 renders the accuracy and usefulness of this interpretation a bit suspect. Thus it might just characterize the extreme ends well while leaving the other observational units more ambiguous.

To get a sense of how accurately the first four principal components represent the data, the quality of representations were evaluated for each observational unit. These are shown in Figure 9. As can be seen in the figure, most observational units are reasonable well represented with 75 percent or more of the quality retained. The importance of the first PC becomes rather obvious, contributing strongly to the representations of many observational units which did not contribute much to the variance of the component. The other components seem less general with contributions to the quality of representations varying from almost 100 percent to non-existent. They characterize specific cases which correspond to the outlying observational units recognized in the original regression analysis. Figure 9 also weakens the interpretation of the third PC since it contributes very little to the representation of Kluuvi which was one of the highest contributing units to the component. This implies that the interpretation is not really relevant when characterizing Kluuvi.

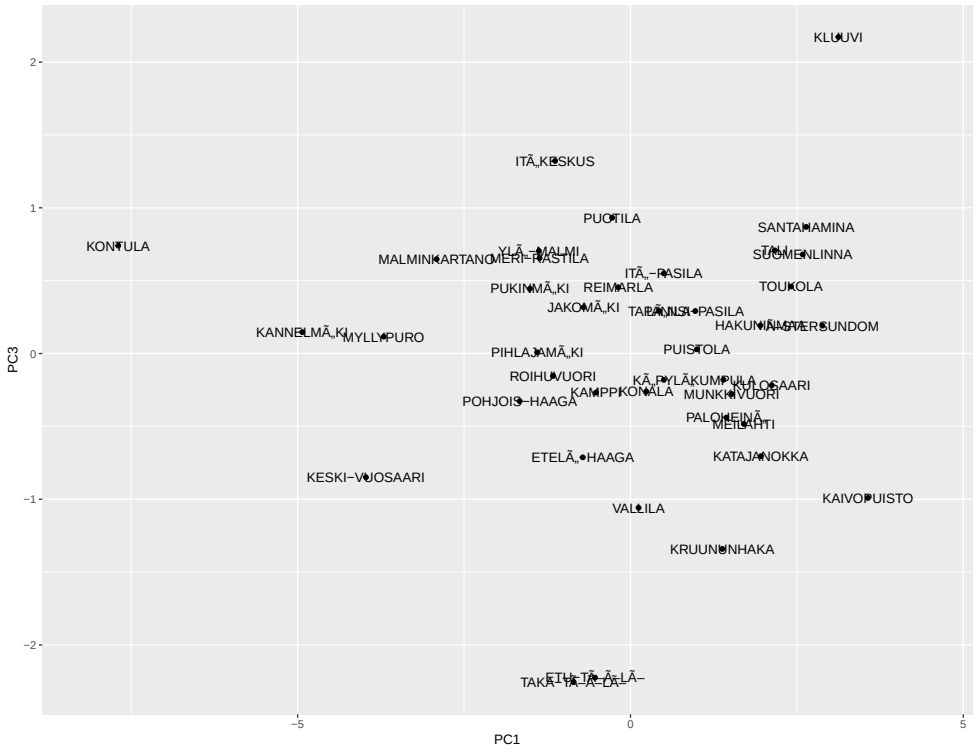

Figure 8: The projection of the observations onto the given principal component subspace.

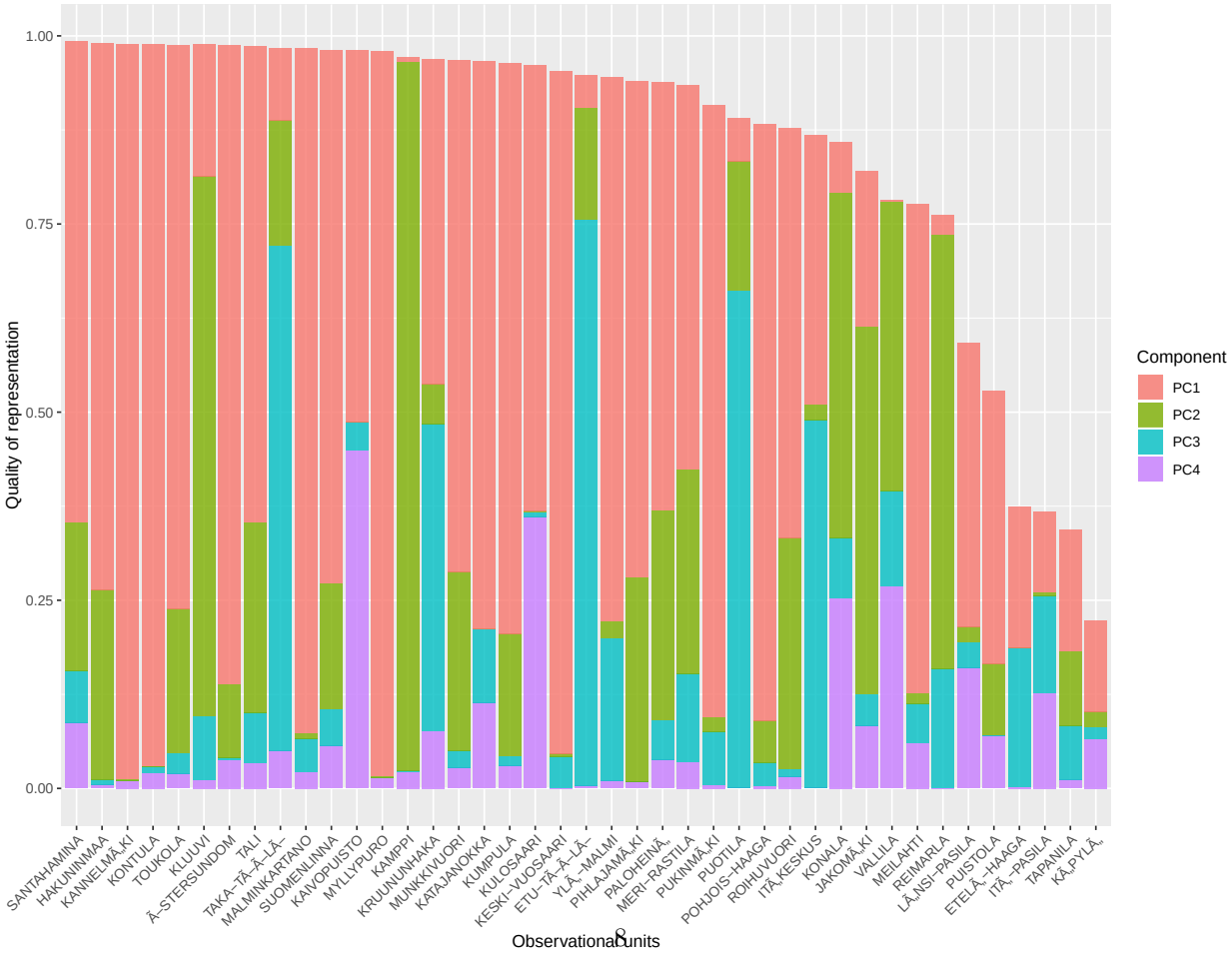

Figure 9: Quality of representation of the observational units by the first four principal components

In order to see how well the principal components model the crime data, the Pearson correlations between them and the crime count variables and the corresponding 95 percent confidence intervals were evaluated. They are shown in Figure 10. It seems that the outlying observational units still effect the estimated associations in a way similar to those shown in Figure 4. This is not surprising since the units were also outlying in terms of crime counts. This masking effect is quite apparent from the scatter plots shown in Figure 11 for the first four principal components. Thus the transformed components retain the same problem that was observed in the original regression problem with Kluuvi and Kamppi being strong outliers. In order to avoid this, one should apply some form of robust regression which is more resistant to these kinds of observations. However, it is still possible to ascertain some associations from the scatter plots. The first component behaves somewhat similarly to the single parent covariate in Figure 5 with most observations behaving quite linearly. The association between PC2 and property crime seems to disperse quite strongly. It is likely that Kluuvi and Kamppi drive the correlation observed in Figure 10. Lastly, there seems to be weak associations between the other considered principal components and property crime. If the interpretation of the third component is the one mentioned earlier, then property crime would tend be more numerous in areas where people flow is related to residents and not to commuters. However, the association could just imply that more residents results in more crime since the population variable had the largest loading on the component.

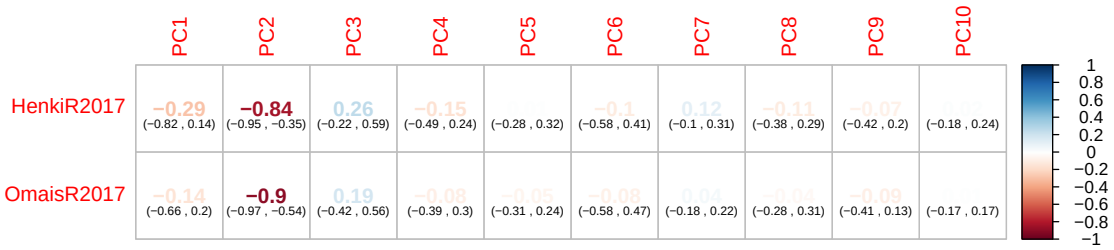

Figure 10: Pearson correlations between the principal components and response variables

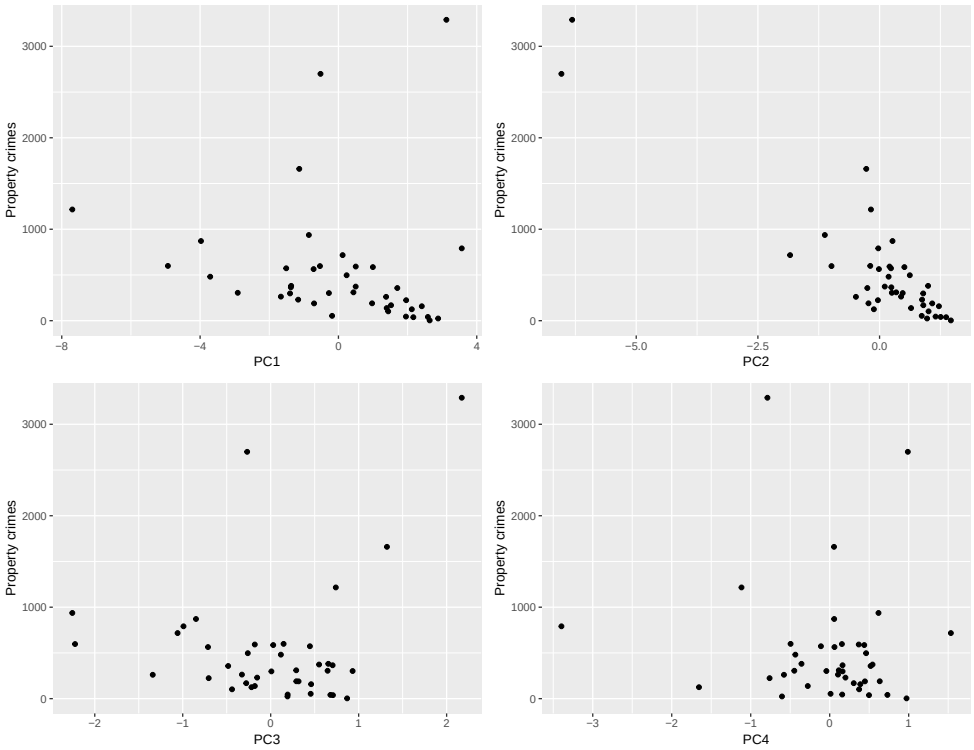

Figure 11: Scatter plots between principal components and property crime

4 Conclusions

The principal component analysis revealed a structure in the covariate data that was in part anticipated from the original regression analysis. The first component was correlated with the socio-demographical variables while the second was formed almost entirely from two observational units where the density of bars and stations is the highest in the region. What was not anticipated was the third component which seemed to solve some of the problems concerning the source of the perpetrators. However, the component contributed very little to the total variance and had small absolute loadings compared to the other components considered in the analysis. In any case, the first four principal components were able to construct 95 percent of the total variance, giving a reasonably sized reduction in the dimensionality of the data with context-relevant interpretations for the components. Of course, it should be considered whether the choice of using non-robust PCA was a reasonable one since it might not be desirable to have a second principal component almost completely defined by two observational units. This problem also appeared in the associations between the principal components and the response variables. Thus one should consider some robust form of PCA and regression analysis to ascertain the associations which are relevant for the majority of the observational units. Even in this case, the regression would most likely be non-linear due to the dependencies observed between the original variables and the responses. It is also not obvious how the principal components would be associated with the responses if some of the original dependencies are nonlinear. This might motivate the use of nonlinear dimensionality reduction methods. All that aside, the project was still able to find meaningful dependencies in the data and provide a good starting point for further analyses.